

UNIVERSIDADE FEDERAL DE PERNAMBUCO
PROGRAMA DE PÓS-GRADUAÇÃO EM
CIÊNCIA DA INFORMAÇÃO
MESTRADO EM CIÊNCIA DA INFORMAÇÃO

REMI CORREIA LAPA

INDEXAÇÃO AUTOMÁTICA NO BRASIL NO ÂMBITO DA
CIÊNCIA DA INFORMAÇÃO
(1973-2012)

Recife
2014



REMI CORREIA LAPA

INDEXAÇÃO AUTOMÁTICA NO BRASIL NO ÂMBITO DA
CIÊNCIA DA INFORMAÇÃO
(1973-2012)

Dissertação apresentada ao Programa de
Pós-Graduação em Ciência da Informação
da Universidade Federal de Pernambuco
como requisito para obtenção do título de
Mestre em Ciência da Informação.

Área de Concentração: INFORMAÇÃO,
MEMÓRIA E TECNOLOGIA.

Linha de Pesquisa: COMUNICAÇÃO E
VISUALIZAÇÃO DA MEMÓRIA.

Orientador: Prof. Dr. Renato Fernandes
Corrêa

Recife
2014

Catálogo na fonte

Bibliotecária Maria Valéria Baltar de Abreu Vasconcelos, CRB4-439

L299i Lapa, Remi Correia

Indexação automática no Brasil no âmbito da Ciência da Informação (1973-2012) / Remi Correia Lapa. - Recife : O Autor, 2014.

287 f.: il.

Orientador: Renato Fernandes Corrêa.

Dissertação (Mestrado) – Universidade Federal de Pernambuco. Centro de Artes e Comunicação. Ciência da Informação, 2014.

Inclui referências e apêndices.

1. Indexação automática - Brasil. 2. Indexação. 3. Recuperação da informação. 4. Sistemas de recuperação da informação. 5. Ciência da Informação. I. Corrêa, Renato Fernandes (Orientador). II. Título.

020 CDD (22.ed.)

UFPE (CAC 2014 – 56)



Serviço Público Federal
Universidade Federal de Pernambuco
Programa de Pós-graduação em Ciência da Informação - PPGCI

REMI CORREIA LAPA

**A INDEXAÇÃO AUTOMÁTICA NO BRASIL NO ÂMBITO DA
CIÊNCIA DA INFORMAÇÃO (1973-2012)**

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de mestre em Ciência da Informação.

Aprovada em: 21/03/2014

BANCA EXAMINADORA

Prof. Dr. Renato Fernandes Corrêa (Orientador)
Universidade Federal de Pernambuco

Prof. Dr. Raimundo Nonato Macedo dos Santos (Examinador Interno)
Universidade Federal de Pernambuco

Profª D^{ca} Mariângela Spotti Lopes Fujita (Examinador Externo)
Universidade Estadual Paulista Júlio Mesquita Filho



Programa de Pós graduação em Ciência da Informação
Av. Reitor Joaquim Amazonas S/N - Cidade Universitária CEP - 50740-570
Recife/PE - Fone/Fax: (81) 2126-7728 / 7727
www.ufpe.br/ppgci - E-mail: ppgciufpe@gmail.com



Aos meus pais, Remi e Jozelia, em
reconhecimento ao que vocês representam
para mim.

Vocês sempre serão meus heróis!

Ao Prof. Dr. Renato Fernandes Corrêa

Um agradecimento especial ao meu estimado orientador, que há cinco anos vem acompanhando o crescimento de nosso trabalho iniciado ainda na Iniciação Científica, pela confiança e ponderações sempre com muita paciência e atenção no processo de orientação.

AGRADECIMENTOS

A Deus;

À minha família pela amizade e apoio;

Aos amigos que sempre estão comigo na presença e na ausência;

À Universidade Federal de Pernambuco;

À CAPES pelo suporte financeiro;

Aos professores e funcionários do PPGCI-UFPE;

À Raissa pelo carinho, apoio e companheirismo;

Aos colegas do PPGCI-UFPE pelo companheirismo durante o curso;

Aos ex-colegas de trabalho da Biblioteca Nilo Peçanha – IFPB, em especial a minha ex-chefe Ivanise por todo apoio. Foi muito bom trabalhar com vocês;

Aos colegas de trabalho da Biblioteca Reitor Ednaldo Bastos – UFPE, pela acolhida;

Às demais pessoas que diretamente ou indiretamente contribuíram na elaboração desta dissertação.

Se esta pequena e modesta contribuição pudesse abrir os horizontes, através de um olhar curioso do que se pensa – e se discute – sobre a informação em outros campos (ou quintais?) da ciência, para uma discussão sobre o alargamento do escopo da ciência da informação, num contexto social, com o uso inteligente das novas tecnologias, na busca de um mundo mais informado, mais consciente e, enfim, melhor, sentir-me-ia bem recompensado.

Jaime Robredo

A única forma de chegar ao impossível é acreditar que é possível.

(Alice no País das Maravilhas)

RESUMO

Apresenta um panorama no âmbito da Ciência da Informação no Brasil referente aos estudos sobre a Indexação Automática por meio do mapeamento e análise da produção acadêmica e científica nacional no período de 1973 a 2012. Como objetivos específicos, procura coletar um *corpus* de análise e caracterizar as pesquisas analisadas quanto aos aspectos metodológicos e de conteúdo; realizar estudos de análise bibliométrica no *corpus* levantado, observando as publicações por autoria, instituições publicadoras, ano, fonte de publicação e instituições acadêmicas; realizar análise de conteúdo observando as publicações pelas categorias: objetivos e aspectos metodológicos. A metodologia consiste em um estudo bibliográfico aprofundado de caráter qualitativo e quantitativo sobre a produção literária no Brasil a respeito da indexação automática de textos escritos no idioma português. O *corpus* de análise para a realização das análises bibliométricas e de conteúdo, é composto de documentos na língua portuguesa, tais como: livros, artigos de periódicos científicos, anais publicados em congressos e seminários, e literaturas cinzentas. Os resultados bibliométricos mostram: a autora Fujita como maior produtora no tema; a revista Ciência da Informação como a que mais publicou sobre a indexação automática; a maior parte da produção se concentra no século XXI; o periódico como a fonte de informação mais utilizada pelos autores como meio de divulgação; a Universidade de Brasília como a instituição acadêmica que mais

produziu trabalhos. Os resultados mais representativos da análise de conteúdo mostram que: 35% dos trabalhos realizaram revisão bibliográfica, em relação aos sistemas/métodos/fórmulas a maioria dos trabalhos de indexação automática, 23%, realizou aplicação e proposição, a proposição corresponde a 20%, e 21% realizaram aplicação; os sistemas como o objeto de estudo mais avaliado, 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, sendo ultrapassada nos estudos dos últimos dez anos pela indexação automática; 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 pesquisa mais utilizado na seleção dos termos. Este estudo propõe ser base de trabalhos futuros na aplicação dos métodos analisados ou na construção de novos métodos de indexação automática.

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

ABSTRACT

Presents an overview of the scope of information science in Brazil related to studies on the Automatic Indexing through the mapping and analysis of national scientific and academic production in the period 1973-2012. Specific objectives , seeks to collect a corpus to analyze and characterize the research analyzed the methodological aspects and content ; conduct studies of bibliometric analysis in the corpus raised, watching the publications by author , publishing institutes , year , source of publication and academic institutions ; perform content analysis by observing the publications categories : goals and methodological aspects . The methodology consists of a detailed bibliographical study of qualitative and quantitative on literary production in Brazil about the automatic indexing of texts written in Portuguese language. The corpus of analysis for the realization of bibliometric and content analysis, consists of documents in the English language, such as books, journal articles, proceedings published in conferences and seminars, and gray literature. The bibliometric results show: the author as Fujita largest producer in the topic; Ciência da Informação magazine as the fastest published on the automatic indexing; most of the production is concentrated in the twenty-first century ; the journal as the main source of information used by the authors as a means of dissemination ; Universidade de Brasília as the academic institution that produced the most jobs . The most significant results of the content analysis show that : 35 % of the work performed

literature review , in relation to the systems / methods / formulas most studies of automatic indexing , 23 % , made application and proposition, the proposition corresponds to 20 % and 21 % had been applied ; systems as the object of more study evaluated and compared to manual indexing as the most used method of evaluation ; the full text as nature 's most researched corpus ; scientific work as the typology of the most studied corpus ; semi-automatic indexing as more applied in terms of validation procedures, and outdated studies in the last ten years by automatic indexing ; the allocation process as the means adopted to identify more terms ; The unstructured text as input data in the preferred systems; natural language as the nature of language , the compound terms like nature of most searched terms ; statistical analysis as the research method most used in the selection of terms . This study aims to be the basis of future work on the application of analysis methods or construction of new methods of automatic indexing .

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

LISTA DE ILUSTRAÇÕES

FIGURA 1 – Modelo simplificado de um sistema de recuperação de informação	45
FIGURA 2 – Rede de relações.	79
FIGURA 3 – Coerência Inter-indexadores.	86
FIGURA 4 – Coerência Intra-indexador.	87
FIGURA 5 – Fatores que podem influenciar a qualidade da indexação.	88
FIGURA 6 – Principais critérios da avaliação.	89
FIGURA 7 – Medida de Revocação.	90
FIGURA 8 – Medida de Precisão.	91
FIGURA 9 – Definições de Revocação e Precisão.	91
FIGURA 10 – Plano horizontal da política de indexação.	95
FIGURA 11 – Plano vertical da política de indexação.	95
FIGURA 12 – Fluxograma simplificado do processo de indexação automática.	107
FIGURA 13 – Fluxograma mais detalhado do processo de indexação automática.	108
FIGURA 14 – Operações sobre o texto.	116

QUADRO 1 – Perguntas que ajudam na identificação do objetivo do documento	63
QUADRO 2 – As etapas da Indexação	68
QUADRO 3 – Comparação entre linguagens de indexação controladas e não-controladas.....	72
QUADRO 4 – Vantagens e desvantagens entre indexação manual e automática	97
QUADRO 5 – Métodos de Indexação Automática	112
QUADRO 6 - Relação entre as instituições acadêmicas	154
QUADRO 7 – Relação Ano e Tipo de validação do termo dos últimos dez anos	188

GRÁFICO 1 – Relação autor x quantidade de pesquisas publicadas	140
GRÁFICO 2 – Relação entre instituição publicadora e frequência de publicação	145
GRÁFICO 3 – Frequência de ocorrência dos trabalhos observados através do Ano de publicação	147
GRÁFICO 4 – Fonte de publicação	149
GRÁFICO 5 – Instituições acadêmicas.....	152
GRÁFICO 6 – Objetivo dos documentos do <i>corpus</i> ..	156

GRÁFICO 7 – Distribuição dos trabalhos que realizaram proposição	163
GRÁFICO 8 – Distribuição dos trabalhos que realizaram aplicação	168
GRÁFICO 9 – Distribuição dos trabalhos que realizaram proposição e aplicação.....	174
GRÁFICO 10 – Frequência de ocorrência dos sistemas nos documentos do <i>corpus</i>	179
GRÁFICO 11 – Frequência de ocorrência dos métodos nos documentos do <i>corpus</i>	180
GRÁFICO 12 – Relação do Nome do Sistema observados nos últimos dez anos.....	181
GRÁFICO 13 – Relação do Nome do Método observados nos últimos dez anos.....	181
GRÁFICO 14 – Distribuição da avaliação dos documentos.....	182
GRÁFICO 15 – Frequência de ocorrência dos métodos de avaliação	183
GRÁFICO 16 – Natureza do corpus.....	184
GRÁFICO 17 – Tipologia do corpus.....	186
GRÁFICO 18 – Validação dos termos	187
GRÁFICO 19 – Identificação dos termos.....	189

GRÁFICO 20 – Frequência de ocorrência da forma como ocorreu a entrada de dados.....	190
GRÁFICO 21 – Natureza da linguagem.....	190
GRÁFICO 22 – Natureza dos termos.....	191
GRÁFICO 23 – Frequência de ocorrência do método/processo de identificação / ponderação / seleção de termos	192
GRÁFICO 24 – Procedimentos internos tomados pelo sistema	193

LISTA DE TABELAS

TABELA 1 – Instituições publicadoras nos últimos dez anos	146
TABELA 2 – Instituições acadêmicas nos últimos dez anos	154
TABELA 3 – Apreciação das categorias utilizadas na análise de conteúdo do corpus em relação ao ano	197

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
ANCIB	Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação
ANO	Ano
ANTE	<i>Abstracts in New Technologies and Engineering</i>
AUT	Autor
BINAGRI	Biblioteca Nacional de Agricultura
BRAPCI	Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação
CBBD	Congresso Brasileiro de Biblioteconomia, Documentação e Ciência da Informação
CDD	Classificação Decimal de Dewey
CDU	Classificação Decimal Universal
CI	Ciência da Informação

Ci. Inf.	Ciência da Informação
CINFORM	Encontro Nacional de Ciência da Informação
CIT	Citação do trabalho
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
ENANCIB	Encontro Nacional de Pesquisa em Ciência da Informação
Enc. Bibli: R. Eletr. Bibliotecon. Ci. Inf.	Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação
Est. Avanç. Bibliotecon. Cienc. Inf.	Estudos Avançados em Biblioteconomia e Ciência da Informação
FAPEMIG	Fundação de Amparo à Pesquisa do estado de Minas Gerais
FONT	Fonte de Publicação
IBBD	Instituto Brasileiro de Bibliografia e Documentação
IBICT	Instituto Brasileiro de Informação em Ciência e

	Tecnologia		
ICT	Informação	científica	e
	tecnológica		
I&S	Informação	&	Sociedade:
	Estudos		
Inf. & Soc.:Est.	Informação	&	Sociedade:
	Estudos		
INST	Instituições		
KWAC	<i>Keyword and context</i>		
KWIC	<i>Keyword in context</i>		
KWOC	<i>Keyword out of context</i>		
LC	<i>Library of Congress</i>		
LD	Linguagem Documentária		
LI	Linguagem de Indexação		
LIPHIS	<i>Linked Phrase Indexing System</i>		
NEPHIS	<i>Nested Phrase Indexing System</i>		
PER	Periódicos		
Perpect. Cienc. Inf.	Perspectivas	em	Ciência da
	Informação		
PLN	Processamento	de	Linguagem
	Natural		

POPSI	<i>Postulated-based Permuted Subject Indexing</i>
PPGCI	Programa de Pós-Graduação em Ciência da Informação
PRECIS	<i>PREserved Context Indexing System</i>
PUC	Pontifícia Universidade Católica
R. Bibliotecon.	Revista de Biblioteconomia
R. Esc. Bibliotecon.	Revista Escola de Biblioteconomia
RBB	Revista de Biblioteconomia de Brasília
RBBd	Revista Brasileira de Biblioteconomia e Documentação
RI	Recuperação da Informação
RS	Rio Grande do Sul
SEBRAE	Serviço Brasileiro de Apoio às Micro e Pequenas Empresas
SIR	Serviço de Indexação e Resumo
SiRILiCo	Sistema de Recuperação de

	Informação baseado em teorias da Linguística Computacional e Ontologia
SISA	<i>Sistema de Indización Semi- automática</i>
SLIC	<i>Selective Listing in Combination</i>
SN	Sintagmas Nominais
SNBU	Seminário Nacional de Bibliotecas Universitárias
SPIRIT	<i>Système Syntaxique et Probabiliste d'Indexation et de Recherche d'Informaticos Textuelles</i>
SRI	Sistema de Recuperação da Informação
SRIAC	Sistema de Recuperação de Informação Assistida por Computadores
TALN	Tratamento Automático da Linguagem Natural
TCC	Trabalho de conclusão de curso
TGT	Teoria Geral da Terminologia

TIC	Tecnologias da Informação e Comunicação
TIT	Título
UFES	Universidade Federal do Espírito Santo
UFGRS	Universidade Federal do Rio Grande do Sul
UFMG	Universidade Federal de Minas Gerais
UFPE	Universidade Federal de Pernambuco
UFRJ	Universidade Federal do Rio de Janeiro
UFScar	Universidade Federal de São Carlos
UFU	Universidade Federal de Uberlândia
UNB	Universidade de Brasília
UNESCO	<i>United Nations Educational, Scientific and Cultural</i> Organization
UNESP	Universidade Estadual Paulista

UNISIST	<i>United Nations International Scientific Information System</i>
USP	Universidade de São Paulo
UTC	Unidades Terminológicas Complexas
WEKA	<i>Waikato Environment for Knowledge Analysis</i>

Sumário

1 INTRODUÇÃO	28
2 FUNDAMENTOS CONCEITUAIS	39
2.1 Sistemas de Recuperação da Informação.....	40
2.2 Indexação	56
2.2.1 Aspectos conceituais da Indexação	61
2.2.2 Linguagens de indexação	70
2.2.3 Avaliação e qualidade na indexação	81
2.3 Indexação Automática	97
2.3.1 Fundamentos teóricos	103
2.3.2 Avaliação da indexação Automática.....	116
3 PROCEDIMENTOS METODOLÓGICOS	122
3.1 Características metodológicas	122
3.2 Considerações sobre o corpus utilizado	125
3.3 Elaboração do panorama	129
3.4 A metodologia prospectiva para a Análise Bibliométrica	131
3.5 A metodologia prospectiva para a Análise de Conteúdo	132
4 APRESENTAÇÃO DOS RESULTADOS E DISCUSSÃO	138

4.1 Análises bibliométricas	138
4.1.1 Autores	139
4.1.2 Instituições publicadoras	143
4.1.3 Ano de publicação.....	147
4.1.4 Fonte de publicação	148
4.1.5 Instituições acadêmicas.....	151
4.2 Análise de conteúdo	155
4.2.1 Análise do objetivo dos documentos do corpus.....	155
4.2.2 Nome Sistema/Método/Fórmula.....	177
4.2.3 Distribuição da avaliação dos documentos	182
4.2.4 Natureza do corpus.....	184
4.2.5 Tipologia do corpus.....	185
4.2.6 Validação dos termos	186
4.2.7 Identificação dos termos.....	188
4.2.8 Forma que ocorreu a entrada dos dados	189
4.2.9 Linguagem de indexação	190
4.2.10 Método/processo de identificação / ponderação / seleção de termos	191
5 CONCLUSÃO.....	199

REFERÊNCIAS	208
APÊNDICE A – Foco de análise para o panorama ordenada cronologicamente	235
APÊNDICE B – Categorias para análise bibliométrica e conteúdo	247
APÊNDICE C – Ficha de análise de conteúdo.....	287

1 INTRODUÇÃO

Os assuntos abordados nesta pesquisa estão circunscritos ao campo da Ciência da Informação (CI), mais especificamente ao tratamento do tema “métodos, técnicas e instrumentos de recuperação da informação”, que é discutido na linha de pesquisa “Comunicação e visualização da memória¹” do Programa de Pós-Graduação da Universidade Federal de Pernambuco (PPGCI-UFPE).

Os avanços promovidos pelas Tecnologias da Informação e Comunicação (TIC) ao longo da história vem elevando o volume da produção da informação exponencialmente. Na argila e na pedra não havia a possibilidade de se escrever longos textos; já com a página tornou-se possível escrever longos textos favorecendo o manuseio, a releitura e a retomada de textos. Mas foi com a invenção da imprensa que houve uma ampliação do “público consumidor” e com a evolução das técnicas de imprimir que houve um

¹ Esta linha de pesquisa se propõe em compreender a comunicação e a visualização da memória como fenômeno da socialização do conhecimento científico mediado pelas tecnologias de informação e comunicação. Neste contexto ela concentra-se nos aspectos práticos e aplicados, contemplados em metodologias e técnicas ligadas à produção, organização, recuperação e disseminação da informação. (UNIVERSIDADE FEDERAL DE PERNAMBUCO, 2009).

incremento das produções científicas e literárias em virtude da rapidez de impressão e circulação das informações (MENDES, 2008).

O estudo da Indexação Automática decorre da necessidade de suprir os efeitos do fenômeno do aumento da produção científica ao propor melhorar o processo de recuperação da informação. 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, 1982a) e (NARUKAWA, 2011).

Neste aspecto a Indexação Automática apresenta características que a liga com algumas das possíveis pesquisas no campo da CI, pois, esta se ocupa com a geração, coleta, organização, interpretação, armazenamento, recuperação, disseminação, transformação e uso da informação, com ênfase particular, na aplicação de tecnologias modernas nestas áreas (CAPURRO; HJØRLAND, 2007).

Esse método automático é conceituado 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.

Os problemas relacionados com a recuperação da informação tornaram-se de interesse para Luhn², que 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).

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 (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).

Alguns estudos se propõem a buscar alternativas eficazes que possibilitem a representação automatizada dos assuntos principais dos documentos (SOUZA, 2006; CÂMARA JÚNIOR, 2007a; MAIA, 2008; BORGES, 2009; NARUKAWA, 2011). Esses estudos nos

² Hans Peter Luhn, especialista da IBM, foi o pioneiro na aplicação da análise estatística de vocabulário para executar uma indexação automática (MORAES, 2002).

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.

Nesse aspecto, a indexação automática ganha um enfoque social, pois auxilia um 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).

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”, como um processo de tratamento temático utilizado na recuperação da informação.

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 este possa ser recuperado quando for solicitado em outro momento (FUJITA, 2009).

De acordo com Araújo Júnior (2007), as instituições demonstram interesse e preocupação em investir em tarefas relacionadas ao processamento da informação, como: indexação, armazenamento e recuperação. Estes três elementos compõem conjuntos de operações realizadas consecutivamente voltadas a localizar, dentro da totalidade de informações disponíveis, aquelas realmente relevantes e por este motivo são considerados por Rowley (2002) como as principais etapas na construção de um Sistema de Recuperação da Informação (SRI). Como já mencionado, o que está despertando cada vez mais interesse na RI é a grande quantidade de informação digital produzida (CANFORA; CERULO, 2004).

Das etapas apresentadas por Rowley (2002), é a indexação que tem o objetivo de representar o documento de forma que este possa ser recuperado posteriormente, portanto, a função do indexador, como afirma Bates³ (1998 *apud* LANCASTER, 2004), é procurar prever os termos que seriam utilizados como descritores na recuperação, o que torna a indexação um processo essencial na realização de uma busca eficiente pelo usuário.

Portanto, é comum o uso de bases de dados, cada vez maiores, por Sistemas de Recuperação de Informação. E cada vez mais comum também é a cena do cientista que pesquisa um banco de dados à procura de uma determinada informação e não obtém todos os documentos pertinentes às suas necessidades informacionais. Esta situação é descrita por Sayão (1996, p. 314) da seguinte forma:

Quando um pesquisador, diante de um microcomputador ligado a um banco de dados que pode estar em qualquer parte do mundo, vasculha suas estantes eletrônicas à procura de informações que definam, completem ou estabeleçam as fronteiras do seu trabalho de pesquisa, ele repete o mesmo gesto de quem mergulha na memória de seu grupo para

³ Bates, M. J. Indexing and Access for digital libraries and the Internet: human, database, and domain factors. **Journal of the American Society for Information Science**, v. 49, n. 13, p. 1185-1205, 1998.

reconstruir as lembranças comuns e dessa forma manter íntegra a sua comunidade. (SAYÃO, 1996, p. 314)

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).

A **problemática** subjacente à dissertação 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) Identificar os autores, períodos, instituições publicadoras, fontes de publicação e instituições acadêmicas, que são mais representativos na Ciência da Informação dentro da temática indexação automática;
- c) Analisar os objetivos, aspectos metodológicos e sistemas, métodos, fórmulas de indexação automática do *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.

Para cumprir o objetivo geral trabalhou-se com os métodos de análise de conteúdo, pois é através desta que os dados coletados nas fontes pesquisadas obtêm, por procedimentos sistemáticos e objetivos de descrição do conteúdo das mensagens, indicadores que permitem a inferência de conhecimentos relativos às condições de produção/recepção (variáveis inferidas) destas mensagens (BARDIN, 2004).

De acordo com Meireles e Cendón (2010), utilizando-se o ano de publicação, constata-se as tendências nas publicações do passado e nas atuais,

podendo ser realizadas previsões de tendências para os próximos anos.

Este estudo desenvolve-se como pesquisa exploratória e se caracteriza quanto aos procedimentos técnicos como pesquisa documental. Para o desenvolvimento do *corpus* constituído por 69 documentos sobre a Indexação Automática no Brasil utilizaram-se livros e capítulos de livros, dissertações e teses, publicações em periódicos e comunicações em anais de congressos e de seminários, destinados a área de Ciência da Informação publicados/comunicados nos últimos 40 anos (1973-2012).

A dissertação é constituída por cinco capítulos e sua estrutura é a seguinte:

- Nesta **introdução** foram apresentados o contexto social da pesquisa, a origem do problema, os objetivos, os pressupostos e a forma como foram encadeadas as temáticas a serem tratadas ao longo do trabalho.
- No segundo capítulo, **Fundamento conceitual**, que compreende o construto teórico deste trabalho, *a priori*, seção 2.1, versa sobre os SRI. Em seguida, são abordados os

fundamentos conceituais, históricos e metodológicos, subdivididos em duas seções: a seção 2.2 aborda a indexação e suas características, enquanto que na seção 2.3 a indexação automática é o assunto abordado.

- O terceiro capítulo, **Procedimentos Metodológicos**, trata da metodologia empregada na pesquisa.
- No quarto capítulo, **Apresentação dos resultados e discussões**, a proposta é mapear e discutir a produção acadêmica sobre a indexação automática no campo da CI dividindo-a em duas subseções, sendo apresentadas na subseção 4.1 as análises bibliométricas, onde se apresentam os resultados e as discussões relacionados à pesquisa bibliométrica. Finalizamos este capítulo proporcionando os resultados e as discussões obtidas através da análise de conteúdo presentes na subseção 4.2.
- No quinto capítulo, **Considerações finais**, são apresentadas considerações sobre os resultados empíricos obtidos a partir da metodologia

empregada no contexto dos fundamentos teóricos e conceituais. Finalmente, são apontadas algumas sugestões de pesquisa.

2 FUNDAMENTOS CONCEITUAIS

Neste capítulo apresentamos o referencial teórico necessário para a compreensão dos estudos e das pesquisas que compõem o corpus de análise desta dissertação.

Dentro do contexto da Organização do Conhecimento, a Indexação é um processo que pode ser associado a Teoria da Complexidade de Edgar Morin, ao partir da idéia de que o processo de indexação deve considerar que o usuário, em seu processo pela busca da informação, trata-se de um indivíduo multifacetado nas esferas psíquicas, sociais, culturais, biológicas e espirituais (MORIN, 2013).

Diante desta visão, o objetivo da indexação perpassa o simples ato de representar o conteúdo dos documentos através dos termos de indexação, existindo uma função maior a que se destinam os termos de indexação que é o de recuperar o documento possibilitando que o usuário tenha acesso à informação nele contida. Por este motivo, iniciaremos tratando dos Sistemas de Recuperação da Informação na subseção 2.1.

Esta ação de indexar um documento pode ser realizada de forma manual, quando cumprida por um

indexador humano, ou automática, quando o processo é desempenhado através de programas de computador. Na sequência serão abordados na subseção 2.2 a Indexação com seus aspectos conceituais, fundamentos lingüísticos e terminológicos, forma de avaliação e políticas de indexação. Em seguida, na subseção 2.3 serão apresentados os fundamentos teóricos da Indexação Automática, assim como os tipos de Sistemas de Indexação Automática e os procedimentos de avaliação.

2.1 Sistemas de Recuperação da Informação

O Sistema de Recuperação da Informação (SRI) pode ser compreendido como um sistema que além de armazenar a informação se encarrega de recuperá-la, esta informação pode estar em textos completos, documentos substitutos (tais como resumos) ou referenciais bibliográficas (CUNHA; CAVALCANTI, 2008).

De acordo com Ferneda (2012) os SRI tem a função de representar o conteúdo dos documentos do *corpus* e apresentá-los ao usuário de uma maneira que lhe permita uma rápida seleção dos itens que satisfazem total ou parcialmente a sua necessidade de informação, formalizada por meio de uma expressão

de busca, portanto a grande quantidade de informação em uma base de dados faz com que os SRI muitas vezes enfrentem dificuldades em recuperar os documentos relevantes à busca realizada pelo usuário.

Quanto a este assunto, Lancaster (2004) menciona que o real problema não se encontra na quantidade de documentos em si, mas em como recuperar o máximo de itens úteis, e o menor número possível de itens inúteis, pois uma das principais dificuldades em encontrar informações relevantes está no fato de haver um elevado número de informação disponível e a maioria irrelevante.

Alguns autores acreditam na possibilidade de melhorar a recuperação em certas situações por meio da indexação, afirma Lancaster (2004). Sob este contexto, Guedes (1994) menciona que nos SRI, a representação precisa do conteúdo temático de documentos é uma condição primordial para a recuperação de documentos relevantes.

Neste caso a indexação assume um papel importante, pois é função do indexador tentar antecipar quais termos seriam utilizados na busca pelas pessoas que possuem lacunas de informação.

Sendo assim, observa-se que o processo de indexação e a Recuperação da Informação (RI) possuem uma relação muito estreita, já que a

indexação tem como finalidade representar o documento de forma que este possa ser recuperado posteriormente. Naves (2004) em seu 'curso de indexação' menciona que a RI é efetuada por meio de um SRI, formado pelos processos de Seleção, Aquisição e Indexação, sendo esta considerada o componente mais importante para a eficácia de um SRI, pois

É a operação central do sistema para a armazenagem e pesquisa das informações. Situa-se, na maioria das vezes, no meio da cadeia documental, no momento da entrada dos documentos no subsistema de armazenamento ou no momento da pesquisa. (GUINCHAT; MENO, 1994, p. 175).

Nessa perspectiva, Lopes (1985) informa que o Serviço de Indexação e Resumo (SIR) deve corresponder às necessidades de informação da biblioteca, uma vez que vão servir de intermediário entre o usuário e o documento original.

Para Rowley (2002), o SRI é composto por três etapas: indexação, armazenamento e recuperação, onde na indexação ocorre o "processo de atribuir termos ou códigos de indexação a um registro ou documento, termos ou códigos esses que serão úteis, posteriormente, na recuperação do documento ou registro" (ROWLEY, 2002).

A atribuição dos termos (palavras-chave) que irão representar os documentos pode ser feita de forma intelectual ou manual, ou seja, realizada por seres humanos que elegem, com base num julgamento subjetivo, termos capazes de representar as informações; ou pode ser feita também automaticamente por computador que seleciona, por meio de um conjunto de instruções programadas previamente, termos mais freqüentes capazes de representar os documentos (ARAÚJO JÚNIOR, 2007).

No armazenamento especifica-se a forma como os SRI guardam suas informações para posteriormente serem recuperadas, e utiliza-se o próprio computador para armazenar tanto os arquivos de documentos como para manutenção das bases de dados (ROWLEY, 2002).

A etapa da recuperação depende muito do sucesso das anteriores. Como afirma Araújo Júnior (2007), "a recuperação da informação se dá pela comparação do que se solicitou com o que está armazenado. Portanto, este processo possui como elementos vitais a indexação e o armazenamento".

Segundo Rowley (2002), a recuperação envolve três etapas:

- Aceitação de uma consulta formulada pelo usuário, representando sua necessidade de informação;
- Execução de uma comparação da consulta com os registros existentes na base de dados, de forma a recuperar aquilo que atenda a necessidade do usuário;
- Produção como resultado a ser fornecido ao usuário, de um conjunto de registros recuperados e que foram identificados com base nessa comparação.

Para que o usuário tenha acesso à informação desejada é importante que ele possua uma boa interação com o sistema. Esta interação ocorre através da consulta ou busca que é a maneira de expressar sua necessidade ao sistema.

A busca pode ser realizada de duas maneiras: através da Recuperação onde o usuário expressa sua necessidade ao sistema em forma de questões ou palavras-chave; ou através da navegação onde o usuário não propõe uma questão ao sistema, mas navega (*browsing*) por categorias e entre os documentos em busca de informações pertinentes.

Para cada um desses modos existem modelos computacionais específicos (FERNEDA, 2012).

A FIGURA 1 permite observar como funciona a representação dos termos do documento e a representação da expressão de busca⁴ em um modelo de SRI, onde no centro do processo se encontra a função de busca⁵.

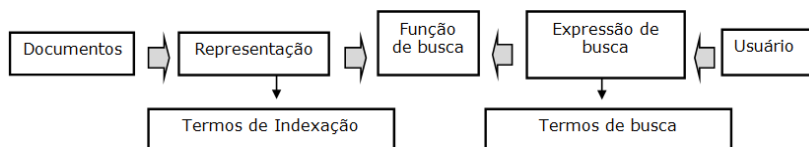


FIGURA 1 – Modelo simplificado de um sistema de recuperação de informação

Fonte: Baseado em Ferneda (2003).

Os SRI exercem sua função de forma satisfatória quando permitem a boa interação do usuário com o sistema permitindo a recuperação de informações relevantes de acordo com a necessidade do usuário, evitando o excesso de informações não desejadas.

Nas palavras de Ferneda (2012) um SRI é a especificação formal de três elementos:

⁴ Expressão de busca é a forma que o usuário emprega para comunicar a sua necessidade informacional para o sistema de informação (FERNEDA, 2012).

⁵ A função de busca compara as representações dos documentos com a representação da expressão de busca e recupera os itens que supostamente fornecerão a informação que o usuário procura (FERNEDA, 2012).

- a) A representação dos documentos;
- b) A representação da necessidade de informação por meio de uma expressão de busca;
- c) E como estes dois elementos serão comparados, a função de busca.

O autor ainda afirma que:

A eficiência de um sistema de recuperação de informação está diretamente ligada ao modelo que ele utiliza, influenciando diretamente em seu modo de operação. Apesar de alguns dos modelos de recuperação de informação terem sido criados nos anos 60 e 70 e aperfeiçoados nos anos 80, as suas principais ideias ainda estão presentes na maioria dos sistemas de recuperação atuais e nos mecanismos de busca da Web. Alguns outros modelos resumem propostas mais recentes que utilizam métodos derivados da Inteligência Artificial e representam alternativas promissoras a serem estudadas e desenvolvidas (FERNEDA, 2012, p. 20, grifo nosso).

Segundo Cardoso (2000) os modelos clássicos utilizados no processo de recuperação de informação são os: booleano, vetorial e probabilístico, que apresentam estratégias de busca de documentos relevantes para uma consulta. Outros modelos mais avançados têm sido propostos ao longo dos anos, dentre estes, destacam-se os modelos baseados em

conhecimento, são os: sistemas especialistas, redes neurais, algoritmos genéticos.

Ferneda (2003) divide os modelos de recuperação da informação em dois grupos:

- *Os Modelos Quantitativos* – são fundamentados pelas disciplinas como a lógica, a estatística e a teoria dos conjuntos. Outra característica comum a estes métodos é que são baseados na semelhança entre a representação dos documentos pelos termos de indexação e pela expressão de busca do usuário. Porém, os termos de indexação associados a um documento não são igualmente úteis para descrever seu conteúdo, pois existem termos mais representativos do assunto principal do documento e outros termos que representam assuntos periféricos à temática do mesmo. Decidir a importância de um termo para a descrição do conteúdo de um documento não é uma tarefa simples, mesmo para pessoas experientes. Alguns sistemas computacionais utilizam propriedades que facilitam a mensuração do potencial representativo de um

termo de indexação. Por exemplo, em um corpus com milhares de documentos uma palavra que aparece em todos os documentos não seria um bom termo de indexação. Por outro lado, uma palavra que aparece em apenas três documentos possivelmente seria de grande utilidade como termo de indexação, pois reduziria consideravelmente o número de documentos que poderiam ser de interesse para uma determinada necessidade de informação do usuário. Portanto, diferentes termos de indexação possuem graus de relevância distintos, de acordo com os documentos e os objetivos do sistema de informação. São exemplos do Modelo Quantitativo: o Modelo Booleano; o Modelo Vetorial; o Modelo Probabilístico; o Modelo Fuzzy; e o Modelo Booleano Estendido.

- *Os Modelos Dinâmicos* – A principal característica dos modelos deste grupo é o reconhecimento da importância do usuário na definição das representações dos documentos. Nesta ótica, os usuários interagem e interferem

diretamente na representação dos documentos do *corpus*, permitindo uma evolução ou uma adaptação dos documentos aos interesses dos usuários do sistema, percebidos através de suas buscas e da atribuição de relevância (e não relevância) aos documentos recuperados (*relevance feedback*). Os sistemas do modelo dinâmico fazem parte de uma classe de sistemas ditos “baseados em conhecimento”, desenvolvidos para servirem como “consultores” na tomada de decisões em áreas restritas. Esses sistemas são adequados para o tratamento de problemas de natureza simbólica, que envolvem incertezas resolvíveis somente com regras de “bom senso” e com raciocínio similar ao humano. Fazem parte deste grupo: os Sistemas Especialistas; as Redes Neurais; e os Algoritmos Genéticos.

Dentre os modelos de recuperação da informação⁶ citados, Ferneda (2003) destaca algumas

⁶ A exposição está bastante simplificada, contendo somente o que se considera necessário para o correto entendimento do que caracteriza os modelos de recuperação da informação citados. Um estudo aprofundado pode ser encontrado em Ferneda (2003).

características e discussões pertinentes a este trabalho, que serão apresentadas a seguir:

- a) *Modelo Booleano* – Neste modelo as buscas são formuladas por meio de uma expressão booleana composta por termos ligados através dos operadores lógicos *AND*, *OR* e *NOT* (respectivamente, E, OU e NÃO), e apresentam como resultado os documentos cuja representação satisfaz às restrições lógicas da expressão de busca. O modelo booleano está presente em quase todos os sistemas de recuperação de informação, seja como a principal maneira de formular as expressões de busca, seja como um recurso alternativo, por duas razões: a primeira está na relativa facilidade de implementação dos algoritmos necessários no desenvolvimento de sistemas booleanos de recuperação de informação; a segunda é a sua flexibilidade e poder, oferecendo certo controle sobre os resultados do sistema.
- b) *Modelo Vetorial* – Também chamado de modelo espaço vetorial, propõe um ambiente no qual é

possível obter documentos que respondem parcialmente a uma expressão de busca. Isto é feito associando-se pesos tanto aos termos de indexação dos documentos como aos termos utilizados na expressão de busca. Esses pesos são utilizados para calcular o grau de similaridade entre a expressão de busca e cada um dos documentos do *corpus*. Como resultado, obtém-se um conjunto de documentos ordenado pelo grau de similaridade de cada documento em relação à busca.

- c) *Modelo Probabilístico* – Neste, dada uma expressão de busca, pode-se dividir o *corpus* (formado por N documentos) em quatro subconjuntos distintos: o conjunto dos documentos relevantes (*Rel*); o conjunto dos documentos recuperados (*Rec*); o conjunto dos documentos relevantes que foram recuperados (*RR*); e o conjunto dos documentos *não relevantes* e *não recuperados*. O conjunto dos documentos *RR* é resultado da interseção dos conjuntos *Rel* e *Rec*. No modelo probabilístico, o resultado ideal de uma busca é o conjunto que

contenha todos e apenas os documentos relevantes para o usuário, isto é, todo o conjunto Rel.

- d) *Modelo Booleano Estendido* – Tenta unir a potencialidade das expressões booleanas com a precisão do modelo vetorial. Por um lado busca-se atribuir maior flexibilidade ao modelo booleano, introduzindo uma gradação no conceito de relevância e, por outro lado, dar maior poder às buscas do modelo vetorial com a utilização dos operadores booleanos. Pode-se dizer que o modelo booleano e o modelo vetorial são casos particulares do modelo booleano estendido. Essa generalização é realizada com a introdução de dois novos parâmetros em relação ao modelo booleano tradicional: os pesos associados aos termos da expressão de busca e os pesos associados aos operadores booleanos.
- e) *Modelo fuzzy* – Busca superar as limitações do modelo booleano generalizando a teoria dos conjuntos tradicional. Na concepção tradicional um elemento pertence ou não a um

determinado conjunto, da mesma forma como no modelo booleano um documento é ou não relevante para uma determinada expressão de busca. Porém, em um sistema de recuperação de informação esse limite claro e preciso não existe. Entre um documento relevante e um não-relevante podem existir gradações (relevância parcial) com as quais o modelo booleano não consegue lidar.

- f) *Processamento da Linguagem Natural (PLN)* – Não se caracteriza como um modelo de recuperação de informação, na medida em que não propõe uma estrutura para a representação dos documentos e não formaliza explicitamente uma função de busca. Porém, o PLN aplicado às expressões de busca de um SRI assume uma importância considerável na medida em que tenta interpretar a necessidade de informação dos usuários. Contudo, essa tarefa é dificultada pelo tamanho (número de palavras) reduzido das expressões de busca que geralmente são utilizadas pelos usuários, não permitindo uma interpretação adequada das expressões.

Portanto, a utilização mais importante do PLN está na interpretação do conteúdo dos documentos a fim de gerar uma representação adequada destes. No entanto, o PLN não elimina a necessidade da utilização de métodos estatísticos e deve ser visto como uma ferramenta complementar aos mesmos.

- g) *Sistemas Especialistas* – Trata-se de um sistema computacional que busca representar o conhecimento de um especialista humano em um domínio particular, de maneira a auxiliar na tomada de decisões e na resolução de problemas relacionados a esse domínio. A ideia subjacente à construção dos sistemas especialistas é que a inteligência não é apenas raciocínio, mas também memória. Pois, é comum considerarmos inteligente uma pessoa que possui grande quantidade de informação sobre um determinado assunto. Assim, os sistemas especialistas obedecem ao princípio de que memória é condição necessária para a inteligência. As idéias relacionadas nestes sistemas podem contribuir para o

desenvolvimento de sistemas que abranjam áreas do conhecimento bastante específicas e em situações nas quais os usuários e os sistemas possam se complementar.

- h) *Redes Neurais* – Modelo computacional que se propõe a simular o funcionamento das células do cérebro. A rede neural procura representar a estrutura de um sistema de recuperação de informação (formado de um lado pelas expressões de busca, do outro lado pelos documentos e no meio pelos termos de indexação); no caso da rede neural essa estrutura seria formada por três camadas: a camada de busca seria a camada de entrada, a camada de documentos seria a saída e a camada de termos de indexação seria uma camada central.
- i) *Algoritmos Genéticos* – Sua aplicação na representação da informação representa um novo modelo para todo o processo de recuperação. As representações dos documentos podem ser vistas como um tipo de “código genético”. Nesse código genético um

cromossomo é representado por um vetor binário onde cada elemento armazena o valor "0" (zero) ou o valor "1" (um), correspondendo respectivamente à presença ou ausência de um determinado termo na representação do documento.

Evidenciamos nesta subseção a relação do processo de indexação com os SRI e a importância de seu emprego na recuperação da informação. Em seguida será dedicada uma atenção especial à indexação, apresentando elementos inerentes a seus conceitos, terminologias, avaliação e política.

2.2 Indexação

Para Robredo (1982a), 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 idéia 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

idéias – pulando, praticamente, as palavras não significativas” (ROBREDO, 1982a, p. 237).

Pode-se nas palavras de Borges (2009) separar o processo de memorização humana em duas etapas principais:

- Memorização temporária e inconsciente – nesta etapa, há a conservação das palavras significativas passando por uma modificação ou aperfeiçoamento a partir da detecção de novos conceitos significativos;
- Memorização permanente dos conceitos assim trabalhados, à qual se atribui o nome de memória.

Segundo Robredo (1982a), no fim do processo encontra-se fixado na memória do leitor uma série de palavras⁷-conceitos⁸-descritores⁹ que representam as idéias básicas do documento por ele lido. A leitura, através de um processo de análise-indexação, leva à

⁷ É um signo linguístico de unidade da língua geral (LARA, 2004).

⁸ É um símbolo mental, uma noção abstrata contida em cada palavra de uma língua que corresponde a um conjunto de características comuns a uma classe de seres, objetos ou entidades abstratas, determinando como as coisas são (LARA, 2004).

⁹ É um signo linguístico qualificado no interior de um discurso de especialidade (LARA, 2004).

armazenagem dos descritores que representam o conteúdo dos documentos.

A aplicação da indexação remonta desde os tempos das tábuas de argila (século II a.C.), onde foram encontradas formas de representação condensada que davam acesso aos conteúdos dos documentos, até o grande desenvolvimento da indexação que ocorre no final do século XIX com o aumento de publicações periódicas e da literatura técnico científica de modo geral (PINTO MOLINA, 1993¹⁰ *apud* FUJITA; RUBI; BOCCATO, 2009).

Gil Leiva e Rodríguez Muñoz aludem que as origens da indexação:

se encontram nas tarefas realizadas pelos antigos escribas da Mesopotâmia. Naquela época, começaram a ter salas para a cópia das tábuas de argila, a elaboração de etiquetas e a conservação das placas. Os textos eram armazenados em prateleiras de madeira, colocados em nichos nas paredes ou eram dispostos em caixas de madeira. Para saber o que continham, colocavam uma pequena etiqueta anexada na lateral, onde escreviam o conteúdo dos documentos. Nessas tarefas rudimentares, vemos os primeiros passos do que hoje conhecemos como a indexação (GIL LEIVA; RODRÍGUEZ MUÑOZ, 1996a, p. 53).

¹⁰ PINTO MOLINA, M. **Análisis documental**: fundamentos y procedimientos. 2.ed. rev. aum. Madrid: Eudema, 1993.

Percebemos que o aumento gradativo da produção de documentos tem reflexo na sociedade, tendo em vista que proporcionam transformações na sua estrutura informacional, que pode ser observado através das modificações ocasionadas no armazenamento, na indexação e na recuperação, elementos que compõem o Sistema de Recuperação da Informação (SRI).

Nessa perspectiva, Anízio e Nascimento (2012) informam que diante da necessidade da organização da informação, para tornar efetivo seu uso e gerar conhecimento, se faz necessário que todo o saber seja representado por meio da indexação, como forma de facilitar a recuperação.

Existe uma dicotomia nas buscas realizadas nos SRI: pode-se recuperar uma grande quantidade de documentos indexando o texto completo ou utilizar a indexação por palavras-chave para diminuir este conjunto de dados correndo o risco de deixar algum documento pertinente fora da pesquisa, desta forma, provocando o silêncio, Quoniam et al. (1998).

Na lição de Chaumier (1988) percebe-se o valor atribuído a indexação quando menciona que ela é a parte mais importante da análise documentária. Conseqüentemente, é ela quem condiciona o mérito de um sistema documentário. O autor nos adverte que

uma indexação inadequada ou uma indexação insuficiente representa 90% das causas essenciais para aparição de 'ruídos'¹¹ ou de 'silêncios'¹² em uma pesquisa.

A indexação é um processo de tratamento temático essencial, pois consiste no ato de identificar e descrever um documento de acordo com o seu assunto, e cujo principal objetivo é orientar o usuário sobre esse conteúdo intelectual, permitindo, dessa forma, a sua recuperação de forma ágil e eficiente (ANÍZIO; NASCIMENTO, 2012).

O propósito da indexação é a representação dos conceitos contidos nos documentos para que se possa realizar a recuperação. Estes conceitos, de acordo com os princípios de indexação elaborados pela *United Nations International Scientific Information System* (UNISIST¹³, 1981) podem ser representados por termos selecionados da linguagem natural ou por símbolos. Para o desenvolvimento deste trabalho

¹¹ São os documentos não pertinentes à questão, que são recuperados em uma pesquisa bibliográfica (CHAUMIER, 1988).

¹² São os documentos pertinentes existentes no acervo, não recuperados durante a pesquisa (CHAUMIER, 1988).

¹³ Programa Internacional que visa à facilitação de Informação Científica criado pela Organização das Nações Unidas para a Educação, a Ciência e a Cultura – UNESCO (SILVA, 1994).

interessa a representação dos conceitos pelos termos de indexação¹⁴.

2.2.1 Aspectos conceituais da Indexação

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).

Na avaliação de Anízio e Nascimento (2012, p. 124) “a representação do conteúdo informacional é uma das etapas mais importantes dentro de uma unidade de informação, pois se um livro estiver mal representado, conseqüentemente, haverá falhas na sua recuperação.”.

O conceito da ‘indexação’ a priori surgiu a partir da elaboração de índices e atualmente está mais ligado à análise de assunto (SILVA; FUJITA, 2004).

¹⁴ Palavra que representa um conceito ou significado presente no documento (MENDES; SIMÕES, 2002).

Para Mai (2005), a determinação do assunto dos documentos, representando-os através de descritores, cabeçalhos de assunto, entre outros, é o que configura como a principal função da indexação, procurando tornar a recuperação do assunto possível.

Portanto, o processo de indexação propõe-se a determinar as palavras-chave, os índices ou os cabeçalhos de assunto de um documento. Segundo Gil Leiva (2012), para obtê-los, previamente foi desencadeado uma sucessão interativa e simultânea dos processos mentais que têm a ver com a percepção da informação, da memória e da compreensão.

Por meio destes processos mentais ocorre a identificação do(s) assunto(s) de que trata o documento através da análise de seu conteúdo para que sejam extraídos os conceitos que irão representá-lo. “Esta representação identifica o documento e define seus pontos de acesso para a busca.” (FERNEDA, 2012, p. 15). Ainda de acordo com Ferneda (2012), essa tarefa, de identificar e descrever os documentos, é geralmente realizada utilizando os procedimentos de indexação.

Na opinião de Guinchat e Menou (1994), a indexação deve responder às questões que os usuários fazem para saber qual é o objetivo de um documento e em que este documento pode servi-lo. Podemos

verificar no QUADRO 1 algumas perguntas formuladas por Guinchat e Menou para ajudar na identificação do objetivo do documento:

QUADRO 1 – Perguntas que ajudam na identificação do objetivo do documento

O que?	Leva a determinar os assuntos ou temas tratados pelo documento ou pela pergunta, como, por exemplo, as técnicas de irrigação, os fertilizantes e as técnicas culturais.
De que forma?	Leva a precisar a forma como os assuntos apresentam-se como, por exemplo, o estudo de uma rede de irrigação e o cálculo das necessidades de água.
Como?	Leva a precisar as diversas circunstâncias que cercam a ação, as causas, as consequências, os objetivos, como, por exemplo, a introdução de novas culturas e a seca persistente, assim como certas formas práticas de realizar uma ação, como a irrigação por aspersão e a utilização de um modelo para estudo.
Quando?	Leva a precisar a data ou o período em que se desenvolve a ação, que é geralmente diferente da data do documento.
Onde?	Leva a precisar o lugar onde se desenvolve a ação, quando se trata de uma zona geográfica determinada. Algumas informações necessitam ser localizadas apenas por país, mas outras devem ser localizadas em escala menor, por estado, cidade, bairro ou região agrícola.

Fonte: Guinchat e Menou (1994, p. 179)

Partindo do mesmo princípio Lancaster (2004) compreende que no processo de indexação devem-se formular várias perguntas sobre o documento, dentre as quais:

- a) De que trata o documento?
- b) Por que foi incorporado ao nosso acervo?
- c) Quais de seus aspectos serão de interesse para nossos usuários?

Na apostila “Oficina de Indexação”, Gaspar (2011) declara a importância de que o indexador

conheça o assunto e que se coloque, na medida do possível, na posição do usuário incorporando outras duas questões às três vistas anteriormente:

- Se eu estivesse interessado neste assunto, como procuraria?
- Se eu estivesse interessado neste assunto, esta informação seria satisfatória?

As respostas a estas questões procuram representar os documentos de forma que os termos da indexação sejam o mais idêntico possível aos termos utilizados pelos usuários na representação da expressão de busca, satisfazendo assim a necessidade informacional dos usuários. Mas segundo Mai (2001) o processo de indexação pode gerar diferentes efeitos interpretativos, que podem ser explicados a partir de estudos dos signos desenvolvidos através dos conceitos semióticos do filósofo Charles Sanders Peirce (1839-1914).

No processo de indexação proposto por Mai (2001) a semiose ilimitada começa através de um signo inicial, o documento. O indexador inicialmente realiza um ato de interpretação, o produto deste primeiro passo da indexação (análise do documento) será um novo (ou segundo) signo, o assunto. Um novo ato de interpretação, o segundo passo, é então

realizado, a fim de converter o que o indexador idealizou sobre o assunto para algo mais gerenciável e concisa para a indexação, o produto deste ato é um novo signo, a descrição de assunto. Por fim, outro ato de interpretação, a terceira etapa, acontece quando se passa a descrição de assunto para uma linguagem de indexação, gerando um novo signo, a entrada de assunto.

Para Fujita (2013) a representação documentária está presente no processo de indexação em dois momentos distintos: na análise de assunto quando se realiza a identificação e a seleção de conceitos mediante representação por conceitos universais tais como a ação, o objeto que sofreu a ação, o agente que praticou a ação e etc, bem como na tradução, etapa na qual os termos serão traduzidos para os termos de uma linguagem documental que compatibilizará tanto os termos indexados quanto os termos que identificam as necessidades dos usuários.

Gil Leiva (1997, p.18-19, tradução nossa) divide estas duas etapas da seguinte maneira “na análise dos documentos e das perguntas para a seleção dos conceitos explícitos ou implícitos; e no armazenamento de conceitos por meio de termos de linguagem natural, ou sua representação em uma linguagem normalizada e controlada”.

Segundo o Unisist (1981) durante a indexação os conceitos são extraídos do documento através de um processo de determinação do assunto, e então traduzidos para os termos de instrumentos de indexação (tais como tesauros, listas de cabeçalhos de assunto, esquemas de classificação, etc.).

Lancaster (2004) argumenta com relação ao processo de indexação que ele é constituído de duas etapas: a análise conceitual, que indica do que trata o documento; e a tradução, que envolve a conversão da análise conceitual de um documento num determinado conjunto de termos de indexação incorporados ao sistema.

Diante do que foi mencionado percebe-se que o processo de indexação pode ser realizado através de duas etapas: a análise e a representação. Entretanto é interessante observar que o número de operações necessárias para a realização do processo de indexação varia de acordo com o autor (MAI, 2000; RUBI; FUJITA, 2003).

A Associação Brasileira de Normas Técnicas (ABNT) através da Norma 12676, de 1992, entende que a indexação compõe-se de três estágios: exame do documento e estabelecimento do assunto de seu conteúdo; identificação dos conceitos presentes no

assunto; tradução desses conceitos nos termos de uma Linguagem de Indexação (LI)¹⁵.

Na visão de Robredo (2005) o processo ocorre em três etapas: análise conceitual do conteúdo do documento; em seguida, a expressão dessa análise por meio de códigos, palavras ou frases representativos do assunto; e finalmente a tradução das descrições dos assuntos para a linguagem de indexação e organização das descrições de acordo com a sintaxe¹⁶ da linguagem de indexação.

Segundo Van Slype¹⁷ (1977 *apud* CHAUMIER, 1988, p. 64), a indexação comporta quatro operações distintas, a saber:

- Conhecimento do conteúdo do documento;
- Escolha dos conceitos a serem representados, baseando-se na aplicação da regra da seletividade e exaustividade;
- Tradução dos conceitos selecionados da forma em que aparecem impressos no documento,

¹⁵ Também conhecida na literatura como Linguagem Documentária (LD).

¹⁶ Entende-se por sintaxe o conjunto de regras que estabelecem as configurações de categorias de palavras consideradas corretas (ANDREEWSKI; RUAS, 1983).

¹⁷ VAN SLYPE, G. **Conception et gestion des systèmes documentaires**. Paris: Ed. d'Organisation, 1977.

para os descritores do thesaurus aplicando a regra da especificidade;

- Incorporação dos elementos sintáticos.

Outros autores ainda indicam cinco ou mais processos. As etapas da indexação e seus respectivos autores podem ser visualizadas no QUADRO 2.

QUADRO 2 – As etapas da Indexação

	ETAPAS	AUTOR/ES
Duas etapas	a) Determinação do assunto; b) Tradução dos conceitos nos termos da linguagem de indexação.	Unisist (1981)
	a) Reconhecimento e extração dos conceitos informativos; b) Tradução desses conceitos na linguagem documental.	Chaumier (1988) ¹⁷
	a) Análise dos documentos e das perguntas para a seleção dos conceitos explícitos ou implícitos; b) Armazenamento das palavras chave tal como estão ou sua normalização por meio de um vocabulário controlado.	Gil Leiva (1997)
	a) Análise conceitual; b) Tradução.	Lancaster (2004)
	a) Análise de assunto; b) Tradução.	Fujita (2013)
	a) Exame do documento e estabelecimento do assunto de seu conteúdo; b) Identificação dos conceitos presentes no assunto; c) Tradução desses conceitos nos termos de uma linguagem de indexação.	Norma 12676 (ABNT, 1992)
Três etapas	a) Análise conceitual do conteúdo do documento; b) Expressão dessa análise por meio de códigos, palavras ou frases representativos do assunto; c) Tradução das descrições dos assuntos para a linguagem de indexação.	Robredo (2005)
	a) Análise do documento; b) Descrição do assunto; c) Entrada de assunto.	Mai (2001)
	a) Contato com o documento; b) Identificação dos conceitos explícitos e implícitos do documento; c) Tradução dos conceitos expressados em linguagem natural por descritores; d) Estabelecimento de ligações sintáticas entre os descritores.	Van Slype (1977)
Quatro etapas	a) Lembrar os objetivos da operação, se necessário; b) Tomar conhecimento prévio do documento; c) Determinar o assunto principal do documento; d) Identificar os elementos do conteúdo que devem ser descritos e extrair os termos correspondentes; e) Verificar a pertinência dos termos selecionados; f) Traduzir os termos da linguagem natural nos termos correspondentes da linguagem documental, se for o caso; g) Verificar a pertinência da descrição; e h) Formalizar a descrição se o sistema prevê regras especiais de apresentação ou de escrita.	Guinchat e Menou (1994, p. 177)
	a) Registros dos dados bibliográficos; b) Análise do conteúdo dos documentos a partir do título, resumo e texto completo; c) Determinação dos assuntos; d) Conversão dos conceitos extraídos em linguagem de indexação; e) Reexaminar a indexação.	Cleveland e Cleveland (1990, p. 104) ¹⁸

Fonte: Fujita, Rubi e Boccato (2009, p. 25) e Gil Leiva (1997, p.197-198)

A representação do assunto de um documento que apresenta 2, 3 ou 4 etapas, dependendo dos diferentes estudiosos, é resumida por Mai (2000, p. 277) da seguinte forma:

O procedimento em duas etapas consiste de uma etapa em que o assunto é determinado e outra etapa em que o assunto é traduzido e expresso na linguagem de indexação. O procedimento de três etapas insere uma etapa na qual o assunto é formulado explicitamente ou implicitamente. No procedimento de quatro etapas, a tradução do assunto para a linguagem de indexação consiste de duas etapas. O indexador primeiro traduz o assunto de seu vocabulário no vocabulário utilizado na linguagem de indexação. Em seguida, o indexador constrói a entrada de assunto na linguagem de indexação na forma de termos de índice, um código de classificação, ou um cabeçalho de assunto.

Segundo Fujita (2003), é na análise de assunto onde ocorre a etapa mais importante da indexação porque identifica e seleciona os conceitos que representam a essência de um documento.

Esta representação do conteúdo dos documentos pode realizar-se através de conjuntos de termos, extraídos diretamente da linguagem natural, permitindo a recuperação da informação, traduzindo-se através das linguagens de indexação, tema que será abordado em seguida.

2.2.2 Linguagens de indexação

No tratamento temático da informação, na opinião de Dias e Naves (2013), é utilizado um instrumento que fornece os termos padronizados para representar o assunto ou assuntos identificados nos documentos analisados, a Linguagem de Indexação (LI).

Na LI considera-se a expressão *linguagem natural* como sinônimo de 'discurso comum', isto é, a linguagem utilizada habitualmente na escrita e na fala, e que é o antagônico de 'vocabulário controlado'. No contexto da recuperação da informação, a expressão normalmente se refere às palavras extraídas dos textos impressos, podendo ser o conteúdo completo de uma publicação (artigo, relatório, ou até um livro) ou parte dela: o resumo, extrato ou apenas o título, e, por isso, considera-se como seu sinônimo a expressão 'texto livre' (LANCASTER, 2004).

Por sua vez, Rivier (1992) e Anízio e Nascimento (2012) relatam que a LI pode ser definida como uma linguagem artificial, ou seja, construída a partir de normas, com o objetivo de propor códigos para o tratamento e recuperação da informação, utilizando a terminologia visando a garantir

referenciais para a organização de campos temático-funcionais.

Na concepção de Fujita (2004), uma LI é um conjunto controlado de termos dotados de regras sintáticas e semânticas, que possui o objetivo de representar os conceitos significativos dos assuntos dos documentos durante a indexação na fase de tradução, e durante a representação do assunto pelo usuário.

Um terceiro tipo de LI é apresentado por Miguéis e Neves (2013), são as linguagens livres de indexação que se designa deste modo por não apresentar qualquer controle dos termos, quer seja formal, quer seja semântico, e cujos pontos de acesso são as palavras ou expressões utilizadas pelos próprios autores no título, resumo e texto integral dos documentos.

Com base nos estudos de Lopes (2002) e Rowley (2002), no QUADRO 3 compara-se as vantagens e desvantagens na adoção das LI não-controladas ou controladas no processo de representação e recuperação da informação.

QUADRO 3 – Comparação entre linguagens de indexação controladas e não-controladas

LINGUAGENS DE INDEXAÇÃO NÃO-CONTROLADA	
VANTAGENS	Permite a imediata representação do assunto no catálogo on-line sem a necessidade de consulta a uma linguagem documentária e a interlocução entre o bibliotecário indexador e usuário pelo fato de utilizarem a mesma linguagem.
	Processo de busca é facilitado com a ausência de treinamentos específicos no uso de uma linguagem documentária.
	Termos de indexação são identificados e selecionados e representados diretamente dos documentos que vão constituir o catálogo on-line. Em ambientes colaborativos (blogs, twitters, e demais redes sociais), a indexação social é realizada pelo usuário na prática da identificação, seleção e representação dos termos a partir dos próprios recursos informacionais.
	Temas específicos citados nos documentos podem ser encontrados.
DESADVANTAGENS	Baixo custo; possível fazer buscas no conteúdo total da base de dados; toda palavra tem valor de recuperação igual; sem erros humanos de indexação; sem demora na incorporação de termos novos.
	O usuário faz um esforço intelectual maior na identificação de termos polissêmicos, sinônimos e homógrafos.
	Alta revocação e baixa precisão na recuperação da informação no catálogo on-line.
	A estratégia de busca deverá ser exaustiva, arrolando todos os conceitos e seus respectivos sinônimos acerca da temática investigativa do usuário.
	Perda de confiança do usuário no catálogo on-line perante um possível resultado negativo.
DESADVANTAGENS	Maior carga de trabalho para o indexador; podem-se perder informações que estejam incluídas implícitas, mas não explicitamente no texto; ausência de vínculos do genérico para o específico; é preciso conhecer o vocabulário da disciplina.
LINGUAGENS DE INDEXAÇÃO CONTROLADA	
VANTAGENS	Controle total do vocabulário utilizado na representação da informação no processo de indexação, amenizando os problemas de comunicação entre bibliotecários indexadores e usuários.
	A partir das notas de escopo do tesouro, os bibliotecários indexadores podem escolher mais adequadamente os descritores que retratam os assuntos dos documentos.
	Uma linguagem documentária bem elaborada e consistente pode ocasionar alta precisão e relevância na recuperação da informação, transmitindo confiança ao usuário perante um possível resultado negativo e, consequentemente, a credibilidade no catálogo on-line.
	As relações hierárquicas, de equivalência e não-hierárquicas dos tesouros auxiliam tanto o bibliotecário indexador, quanto o usuário na identificação de conceitos relacionados e referência de termos associados.
	Resolve muitos problemas semânticos; permite que relações de gênero-espécie sejam identificadas; mapeia áreas do conhecimento.
DESADVANTAGENS	Necessidade de disponibilização da linguagem documentária na interface de recuperação da informação para o usuário final.
	Necessidade de treinamento no uso do tesouro, tanto para usuários profissionais, quanto para usuários finais.
	Desatualização da linguagem documentária poderá conduzir a representação de assuntos inadequados e resultados insatisfatórios e negativos ao usuário.
	Custo na construção e na gestão, além da necessidade constante de atualização e manutenção da linguagem documentária.
	Custo elevado; possíveis inadequações de cobertura; erro humano; possibilidade de vocabulário desatualizado; dificuldade de incorporar sistematicamente todas as relações relevantes entre os termos.

Fonte: Adaptado de Lopes (2002, p. 47-48) e Rowley (2002, p. 171)

Na opinião de Pinto (1985) a adoção do procedimento das linguagens de indexação não-controladas na recuperação de informação pode causar vários problemas, como a dispersão terminológica¹⁸ e a dispersão sintática¹⁹, pois a linguagem de indexação deve ser compatível com a linguagem de busca do usuário. Sobre este assunto Boccato (2009, p. 21) comenta que “quando isso não ocorre, o sistema de recuperação da informação fica comprometido, ocasionando sérios problemas na recuperação e gerando a insatisfação do usuário”.

Portanto, a linguagem de indexação controlada deve atuar nos sistemas de informação tanto para orientar o indexador sobre quais são os termos mais adequados para representar o assunto de um documento, quanto para guiar os pesquisadores sobre o modo de escolher os termos indexados que representam no sistema o assunto procurado. Tem ainda como função recuperar documentos com conteúdo semelhante, recuperar documentos relevantes sobre um assunto específico, recuperar

¹⁸ Quando diferentes autores usam palavras diversas para expressar o mesmo conceito (PINTO, 1985).

¹⁹ Quando se usam diferentes estruturas para expressar a mesma ideia (PINTO, 1985).

documentos por grandes áreas de assunto, auxiliar na escolha do termos adequado para a estratégia de busca, representar o assunto de maneira consistente permitindo a compatibilidade e o diálogo entre a linguagem do autor, do indexador e a do pesquisador (MOURA et al., 2005).

Por essas razões, a maioria dos SRI adota algum tipo de controle sobre os termos usados onde segundo Rowley (2002) foram demonstrados através de alguns exemplos de buscas realizados em catálogos em linha de acesso público que as linguagens controladas na indexação são mais sistemáticas e, portanto, mais eficientes e de fácil compreensão para quem faz a busca.

Outro benefício na adoção dos processos de LI controladas está no controle de problemas relacionados a questões lingüísticas como a polissemia²⁰ e a sinonímia²¹.

Nesse sentido, o vocabulário controlado é usado para melhorar a eficácia do armazenamento de informação e dos sistemas de recuperação aos que procuram identificar e encontrar o assunto desejado

²⁰ Multiplicidade de significados para um significante (palavra) (ACADEMIA BRASILEIRA DE LETRAS, 2008).

²¹ Equivalência entre os significados de duas ou mais palavras (ACADEMIA BRASILEIRA DE LETRAS, 2008).

por meio da descrição de assunto, usando uma determinada língua. (BOCCATO, 2012).

Dias e Naves (2013) dividem a linguagem de indexação controlada da seguinte forma:

Existem dois tipos principais de linguagens de indexação: linguagens alfabéticas e linguagens simbólicas. Os principais tipos de linguagens alfabéticas que existem são as listas de cabeçalhos de assuntos e os tesauros. Os cabeçalhos são para uso, principalmente, de bibliotecas públicas e outros tipos de bibliotecas cuja abrangência de assuntos seja mais geral. Os tesauros, por outro lado, são voltados para coleções especializadas, seja em determinado campo do conhecimento, seja em qualquer área orientada para uma missão. Quanto às linguagens simbólicas, estas se limitam praticamente aos sistemas de classificação bibliográfica, que utilizam símbolos, como letras e números para representar os assuntos dos documentos (DIAS; NAVES, 2013, p. 13-14).

As linguagens de indexação controlada e de recuperação da informação podem ser pré-coordenadas ou pós-coordenadas, dependendo da necessidade do sistema de informação ou do centro de documentação onde são utilizadas. As linguagens pré-coordenadas são aquelas nas quais os termos que as integram se coordenam num processo que precede sua utilização, enquanto nas pós-coordenadas os termos que as integram se coordenam num processo

posterior a sua fixação, por exemplo no momento de seu estabelecimento ou de seu uso (CURRÁS, 2010).

Como exemplo das linguagens pré-coordenadas podemos citar a Classificação Decimal Universal, a Classificação Decimal de Dewey ou as classificações facetadas, como a *Colon Classification* de Ranganathan.

Sobre os dois tipos principais de linguagens pós-coordenadas Araújo Júnior (2007, p. 36-37) define:

- a) Lista alfabética de termos autorizados – é constituída por uma coleção de conceitos destinados a representar, de maneira unívoca, o conteúdo dos documentos e das buscas em um dado sistema. Tais conceitos são expressos por palavras ou por expressões extraídas de uma lista finita estabelecida a priori. Somente os termos que figuram nesta lista podem ser empregados para indexar os documentos e as buscas.

Exemplo:

Ao se definir LIXO DOMICILIAR como termo autorizado, o sistema não permite que o indexador utilize outros termos, tais como 'resíduo domiciliar' ou 'entulho domiciliar'.

- b) Tesauros – lista estruturada de conceitos destinados a representar, de maneira unívoca, o conteúdo dos documentos e das buscas em um dado sistema, e a apoiar o usuário na indexação dos documentos. Os conceitos são extraídos de uma lista finita, estabelecida a priori. Somente os termos que figuram nesta lista podem ser empregados para indexar os documentos e as buscas. Possui uma estrutura semântica com relações entre os termos de equivalência, hierarquia e associação.

Exemplo:

BIBLIOTECONOMIA

UP Documentação

TG Ciência da Informação

TE Bibliotecário

TA Arquivística

Onde, UP (Usado Para) representa o termo equivalente (mas que não é considerado como descritor); TG representa o Termo Genérico; TE

representa o Termo Específico; e TA o Termo Associado.

Segundo Lancaster (2004, p. 19), a estrutura semântica deste tipo de instrumento de indexação destina-se a:

- a) Controlar sinônimos, optando por uma única forma padronizada, com remissivas de todas as outras;
- b) Diferenciar homógrafos. Por exemplo: PERU (País) é um termo bastante diferente de PERU (ave); e
- c) Reunir ou ligar termos cujos significados apresentem uma relação mais estreita entre si.

Desta forma, nas linguagens pós-coordenadas, utilizadas em sistemas de indexação pós-coordenados, a combinação dos termos ocorre no momento da recuperação da informação e o seu conteúdo temático, que é o objeto de um texto informacional, será também representado pelos termos de indexação que lhe são atribuídos, revestindo-se de um caráter multidimensional, ilustrado na FIGURA 2. Os termos são apresentados hierarquicamente, mostrando as categorias de assunto e relações entre conceitos (ANÍZIO; NASCIMENTO, 2012).

Lancaster (2004, p. 274) afirma que “um sistema baseado em vocabulário pós-coordenado oferece todas as vantagens da linguagem natural e muitos dos atributos do vocabulário pré-controlado”. O autor também argumenta que os sistemas pós-coordenados são sistemas informatizados modernos, funcionando em linha e que podem ser visto como descendentes diretos dos sistemas manuais de ficha.

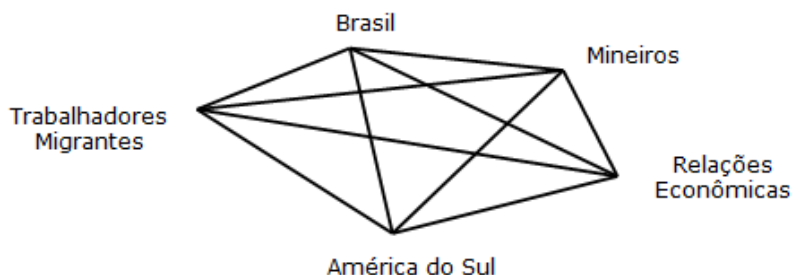


FIGURA 2 – Rede de relações.
Fonte: Baseado em Lancaster (2004).

De acordo com Pinto (1985), qualquer conjunto de termos e/ou símbolos usados para denominar classes de assuntos em um SRI constitui uma LI. E assim, como qualquer língua, uma linguagem de indexação é formada por dois componentes básicos: o vocabulário e a sintaxe. O vocabulário se refere à relação dos descritores usados para a identificação do conteúdo de um documento e a sintaxe se refere às regras empregadas para a combinação dos descritores

usados para a identificação do conteúdo de um documento. (FUJITA, 2004).

Nesse sentido o vocabulário é um instrumento de controle terminológico. Este instrumento permite traduzir a linguagem natural dos documentos, dos usuários e dos indexadores em uma linguagem sistemática de recuperação da informação. (BRASIL, 2002).

Na concepção de Fujita, Boccato e Rubi (2010), a evolução da principal lista de cabeçalhos de assunto, a *Library of Congress Subject Headings* (LCSH) em formato de um tesouro, considerada linguagem de indexação, é o exemplo mais característico de que existe uma tendência mundial de pensar a catalogação de assuntos com o rigor científico e metodológico da indexação para obter especificidade e precisão tanto na análise e representação de assuntos quanto na recuperação, pois os cabeçalhos de assunto, anteriormente pré-coordenados por catalogadores não são necessários após a evolução dos catálogos manuais em catálogos on-line que permitem buscas por qualquer palavra do assunto coordenado com outras palavras ou não.

Por tudo que foi visto, a Linguagem de Indexação, enquanto instrumento de representação do assunto do documento, deve ser avaliada e ter sua qualidade medida, tema abordado a seguir.

2.2.3 Avaliação e qualidade na indexação

O termo 'avaliação' é definido como o 'ato de determinar ou medir o valor' de uma atividade ou objeto (ACADEMIA BRASILEIRA DE LETRAS, 2008).

De acordo com Lancaster (1996), a avaliação é o único caminho que leva à melhoria dos serviços e produtos oferecidos aos usuários e à detecção dos pontos positivos e negativos da usabilidade de um catálogo em linha ou em fichas. O autor ainda apresenta algumas prováveis razões para a realização da avaliação:

- Estabelecer uma espécie de 'escala' para mostrar em que nível de desempenho o serviço está funcionando no momento;
- Comparar o desempenho de várias bibliotecas ou serviços;
- Justificar a existência do serviço ou produto;

- Identificar as possíveis causas de malogro ou ineficiência do serviço, visando a elevar o nível de desempenho no futuro.

Quanto aos métodos de avaliação:

Podem ser subjetiva ou objetiva. Estudos subjetivos, baseados em opiniões, não deixam de ser úteis, pois é importante saber o que as pessoas sentem em relação ao serviço. Mas a avaliação terá sua utilidade máxima se for analítica e diagnóstica, procurando descobrir como o serviço poderia melhorar, e é difícil basear este tipo de estudo apenas em opiniões. Em geral, portanto, devem-se adotar critérios e procedimentos objetivos. Os resultados de um estudo objetivo devem ser quantificáveis (LANCASTER, 1996, p. 10).

Já foi dito anteriormente que “a indexação não constitui um fim em si mesma” (LANCASTER, 2004, p. 83), uma vez que “a indexação não tem razão de existir se não for para recuperar a informação” (GIL LEIVA, 2012, p. 96).

Neste sentido, uma indexação de qualidade não é mensurável através da quantidade de descritores representativos de determinado documento, mas, sim, a que durante as buscas se recuperem o maior número de documentos úteis com o menor número de ruídos e/ou silêncio.

Sobre a qualidade da indexação, Gil Leiva (2012) relata que os elementos que a caracterizam são: a exaustividade, a consistência, a especificidade e a correção.

A respeito da exaustividade, Slype²² (1991 *apud* ARAÚJO JÚNIOR, 2007) expõe que esta mede a qualidade na seleção dos conceitos de fatos significativos, isto é, que contém informação pertinente para os usuários:

- Exaustividade muito reduzida ocasionará a não recuperação de documentos pertinentes, por exemplo, ao realizar uma pesquisa utilizando a expressão de busca 'Clonagem', documentos relevantes que abordem assuntos correlacionados como 'genoma' e 'mapeamento genético' podem ser deixados de fora do resultado de busca; e
- Exaustividade muito elevada ocasionará a recuperação de documentos que não contenham informação pertinente sobre os conceitos da consulta, por exemplo, se a expressão de busca

²² SLYPE, G. V. **Los lenguajes de indización:** concepción, construcción y utilización em los sistemas documentales. Madrid: Fundación Germán Sánchez Ruipérez, 1991.

for 'engenharia genética', mas o interesse está apenas em recuperar documentos sobre 'organismos transgênicos', documentos sobre clonagem animal possivelmente seriam apresentados representando ruído no processo de recuperação da informação.

A Unisist (1981) indica que a exaustividade na indexação propõe identificar todos os conceitos de um documento que teriam um valor potencial para os usuários de um sistema de informação.

Referente à especificidade, Araújo Júnior (2007, p. 44) diz que esta mede a qualidade na seleção dos descritores que correspondem aos conceitos inclusos no documento.

Como regra, na especificidade, os conceitos devem ser identificados tão especificamente quanto possível, evitando um cabeçalho mais geral. Conceitos mais genéricos podem ser selecionados em algumas circunstâncias, dependendo dos objetivos do sistema de informações, por exemplo, em certos casos o indexador pode considerar que uma idéia não foi completamente desenvolvida, ou foi tratada de forma superficial pelo autor, justificando-se a indexação num nível mais geral (UNISIST, 1981).

Quanto à correção na indexação ou a ausência de erros, Soergel (1994) argumenta que esta qualidade tem uma importância primordial para o desempenho da recuperação, pois a indexação é suscetível a dois tipos de falhas²³: por omissão (quando um termo que deve ser atribuído é omitido) e por inclusão (adição de um termo que não deve ser atribuído). A fim de determinar esse tipo de erro é preciso saber quais os descritores que devem e os que não devem ser atribuídos, para isso a correção da indexação pode ser determinada através de um consenso entre vários bons indexadores e usuários experientes.

A consistência ou a coerência na indexação foi definida por Gil Leiva (1997, p. 22), como "o grau de concordância no momento da representação da informação essencial de um documento por meio de um conjunto de termos de indexação selecionados por dois sistemas diferentes ou por cada um dos indexadores de um grupo".

A consistência ou a coerência na indexação é de acordo com Strehl (1998) um dos quesitos para uma indexação de qualidade, já que todos os documentos que tratam sobre o mesmo assunto devem estar

²³ Desvios das regras de indexação (SOERGEL, 1994).

representados da mesma forma, não havendo perda de informações no momento da recuperação.

Portanto, é fundamental que se alcance uma coerência na indexação, entretanto, por ser o ato de indexar um processo subjetivo, “a indexação envolve julgamento, e conseqüentemente, oscila muito no seu nível de concordância e apresenta discrepâncias” (PINHEIRO, 1978, p. 109).

O que dificulta a coerência na indexação, segundo Strehl (1998), é que distintos indexadores (ou até o mesmo indexador em momentos distintos) percebem de forma diferente os descritores definidos para representar os conceitos do conteúdo do documento.

Segundo Lancaster (2004), quando houver concordância entre os diferentes indexadores, ocorre à *coerência inter-indexadores* (FIGURA 3); e no momento em que a concordância for entre o mesmo indexador em diferentes momentos ocorre à *coerência intra-indexador* (FIGURA 4).

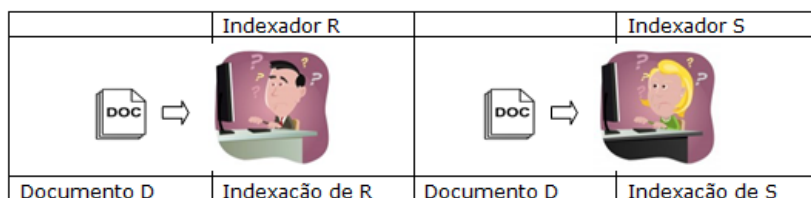


FIGURA 3 – Coerência Inter-indexadores.

Fonte: Baseado em Gil Leiva (2012)

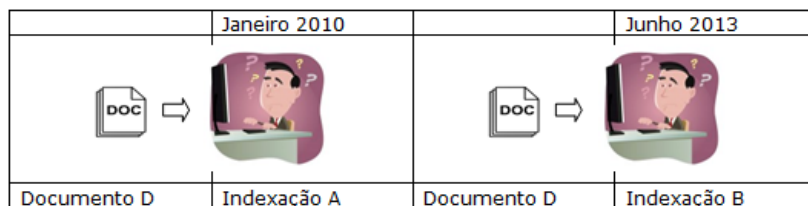


FIGURA 4 – Coerência Intra-indexador.

Fonte: Baseado em Gil Leiva (2012)

Entretanto a coerência na indexação não significa que os descritores utilizados foram os mais apropriados, pois:

analisando-se a coerência da indexação isoladamente, sem considerar o nível de eficácia da recuperação de informações no sistema pelo usuário, não se pode inferir sobre a qualidade da indexação. A coerência, portanto, não significa necessariamente que os documentos foram adequadamente representados, tornando-se com isso muito difícil estabelecer os meios para medir uma 'boa indexação'. (STREHL, 1998, p. 330).

Sendo assim, para que se tenha qualidade na indexação é imprescindível que ocorra coerência inter e intra-indexadores. Por outro lado, apenas a coerência na indexação não denota qualidade na indexação.

Nesse contexto, além dos fatores ligados ao conhecimento do indexador sobre o assunto que podem levar a indexação de termos excessivos,

Lancaster (2004) identifica outros fatores que podem afetar a qualidade da indexação, FIGURA 5.

No que concerne aos **Aspectos da avaliação**, a indexação não é um processo que deve ser considerado como fim em si mesmo, portanto, seus resultados devem ser avaliados e isso somente pode ser feito no contexto de determinada base de dados, seja ela em formato impresso ou eletrônico. Nesse contexto, a indexação é avaliada como bem-sucedida quando permite a quem realiza as buscas localizar itens de que precisa sem ter de examinar muitos de que não precisa (LANCASTER, 2004).

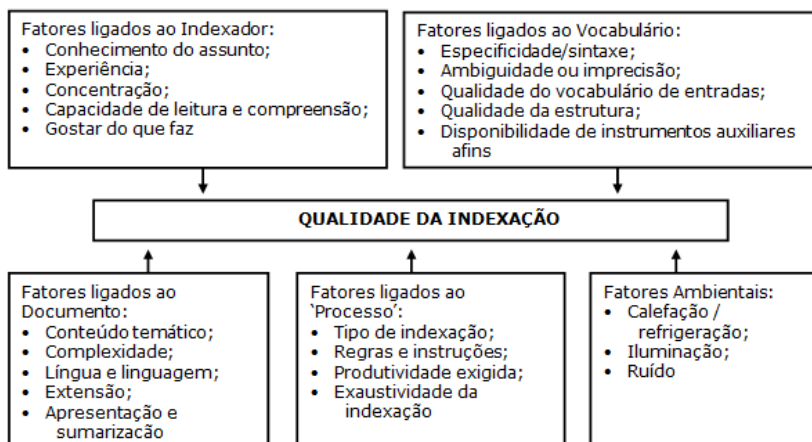


FIGURA 5 – Fatores que podem influenciar a qualidade da indexação.

Fonte: Baseado em Lancaster (2004)

Esta capacidade do sistema em fornecer respostas que realmente correspondam à questão

proposta – a relevância, é possível também, graças à avaliação do SIR que envolve um julgamento do valor da informação recuperada, conforme o entendimento proposto por Lopes (1985).

Para Lancaster (2004, p. 135) “uma base de dados bibliográficos não pode ser avaliada de forma isolada, mas somente em função de sua utilidade ao responder a várias necessidades de informações”.

Existem dois processos de avaliação da indexação: avaliar pela recuperação e avaliar pela linguagem. No primeiro processo a avaliação observará através de medidas definidas os resultados apresentados na recuperação da informação de determinado documento; no segundo processo de avaliação serão analisados se os termos recuperados satisfazem o processo de busca.

Nesse sentido, no que concerne a determinada necessidade de informação, Lancaster (2004) apresenta quatro critérios essenciais, apontados na FIGURA 6.

Cobertura – Quantos documentos sobre um assunto, publicados durante determinado período, se acham incluídos na base de dados? Recuperabilidade – Quantos documentos sobre o assunto, incluídos na base de dados, são encontrados com o emprego de estratégias de busca 'razoáveis'? Previsibilidade – Ao utilizar informações da base de dados, com que eficiência o usuário pode aferir quais os itens que serão e os que não serão úteis? Atualidade – Os itens publicados recentemente são recuperáveis, ou atrasos na indexação provocam uma situação em que os itens recuperados mostram resultados de pesquisas 'antigos' ao invés de 'novos'?
--

FIGURA 6 – Principais critérios da avaliação.
Fonte: Baseado em Lancaster (2004)

Naves (2004) apresenta duas medidas que possibilitam a avaliação de uma Linguagem de Indexação (LI) para um SRI em uma base de dados: a revocação e a precisão.

De acordo com Lopes (1985) a revocação é a capacidade do SIR de oferecer, em resposta a uma questão, todas as referências relevantes existentes na base de dados. E é medida pela proporção de referências relevantes recuperadas, em relação ao total de referências relevantes existentes no SIR, conforme mostrado na FIGURA 7.

$$\text{Revocação} = \frac{\text{Nº de referências relevantes recuperadas}}{\text{Nº total de referências relevantes existentes no SIR}} \times 100$$

FIGURA 7 – Medida de Revocação.

Fonte: Baseado em Lopes (1985)

A Precisão pode ser conceituada como:

a capacidade do SIR em fornecer apenas referências relevantes, eliminando as que não são relevantes para a questão. É uma medida importante para o usuário porque quanto maior a precisão, menor será o esforço do usuário em utilizar o SIR, isto é, os ruídos são eliminados. (LOPES, 1985, p. 249-250).

Segundo Naves (2004), a precisão é medida pelo coeficiente entre o número de documentos relevantes recuperados, em relação ao número total

de documentos recuperados. A fórmula pode ser vista na FIGURA 8.

$\text{Precisão} = \frac{\text{Nº de referências relevantes recuperadas}}{\text{Nº total de referências recuperadas}} \times 100$

FIGURA 8 – Medida de Precisão.
Fonte: Baseado em Lopes (1985)

Segue na FIGURA 9, uma ilustração idealizada por Câmara Júnior (2007b) que resume os conceitos abordados anteriormente relativos à relevância, precisão e revocação. Uma base de documentos e os extratos utilizados para construção das medidas de avaliação para sistemas de recuperação de informação são delineados.

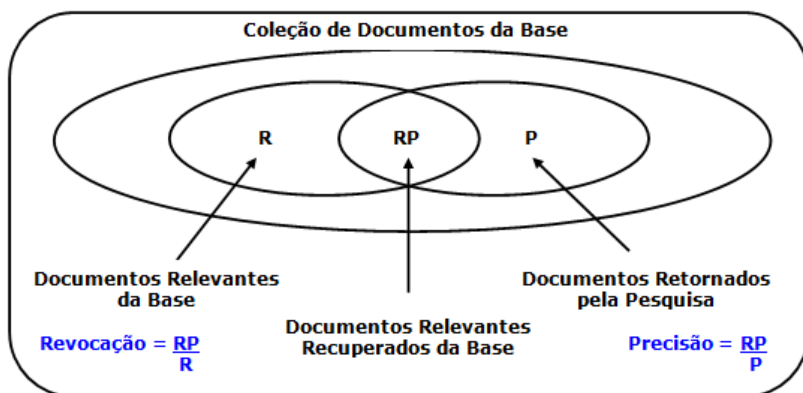


FIGURA 9 – Definições de Revocação e Precisão.
Fonte: Baseado em Câmara Júnior (2007b)

Como visto anteriormente, o objetivo de um serviço de informação é o de satisfazer as

necessidades informacionais do usuário. Para tanto é imprescindível que o processo de indexação seja eficiente em satisfazer as buscas realizadas pelos usuários, além de incluir critérios e medidas referentes à qualidade dos termos indexados, diminuindo assim o esforço e o tempo despendidos pelo usuário na obtenção da informação desejada (LOPES, 1985).

Neste sentido, para que se obtenha um SRI é preciso que seja estabelecida uma política de indexação, que será o tema abordado na próxima seção.

2.2.4 Políticas de indexação

De acordo com Lancaster (2004) a política de indexação é um fator que influencia no desempenho de um SRI e que está diretamente atribuída à complexa atividade que é o processo de indexação.

Por este motivo a política de indexação deve levar em consideração "todo o contexto em que está inserido o sistema de informação: a organização a que serve, a clientela a que se destina e os recursos financeiros, materiais e humanos disponíveis" (STREHL, 1998, p. 330).

Compartilhando desta visão, Fujita (2012) faz um alerta para que a política de indexação não seja

vista apenas como uma série de procedimentos que devam ser seguidos, e sim um conjunto de decisões que esclareçam os interesses e princípios de um sistema de informação e, particularmente, do sistema de recuperação da informação.

Neste sentido, são objetivos de uma política de indexação definir as variáveis que irão afetar consideravelmente o desempenho do serviço de indexar, estabelecer princípios e critérios que servirão de guia na tomada de decisões para a otimização do serviço, racionalização dos processos e consistência das operações nele envolvidas (CARNEIRO, 1985).

Na opinião de Carneiro (1985), alguns fatores que são imprescindíveis ao planejamento de qualquer SRI devem ser levados em conta na elaboração de uma política de indexação, tais como:

- a identificação das características e objetivos da organização à qual estará vinculado o sistema de indexação;
- a identificação do usuário a que se destina o sistema;
- os recursos humanos, materiais e financeiros.

Segundo Strehl (1998), entre os elementos que compõem uma política de indexação podem-se destacar os seguintes:

- a) Cobertura de assuntos – identificação das áreas que necessitam de um tratamento aprofundado e das áreas a serem superficialmente tratadas;
- b) Processo de indexação – definição das variáveis que se referem aos níveis de exaustividade e especificidade requeridos pelo sistema, linguagem de indexação, capacidade de revocação e precisão do sistema;
- c) Estratégia de busca – definição da responsabilidade para realização da busca de informações em um sistema, ponto em que se decide se o bibliotecário ou usuário acessará diretamente a base de dados;
- d) Tempo de resposta do sistema – identificação do tempo permitido para ser consumido no momento da recuperação de informações úteis, sendo determinado pelas exigências de revocação e precisão de um sistema;
- e) Forma de saída – definição da forma de apresentação das informações recuperadas no sistema;
- f) Avaliação do sistema – identificação da forma como o sistema será avaliado visando a

descobrir o nível de satisfação das necessidades de seus usuários, as falhas que estão ocorrendo e a forma como poderão ser corrigidas.

Fujita (2012) em seu livro sobre política de indexação apresenta duas visões que se complementam: uma visão mais ampla, que considera o contexto da gestão de sistemas de recuperação da informação sobre as atividades de indexação, neste caso a política de indexação estrutura-se a partir do eixo horizontal (FIGURA 10) das atividades de gestão, e a visão mais direcionada aos procedimentos e elementos de indexação relacionados ao manual de indexação, sua operacionalização e avaliação, correlacionado ao eixo vertical (FIGURA 11) de uma política de indexação.

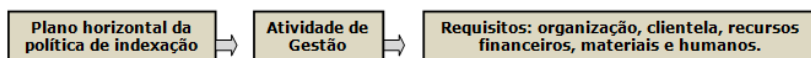


FIGURA 10 – Plano horizontal da política de indexação.

Fonte: Baseado em Fujita (2012)

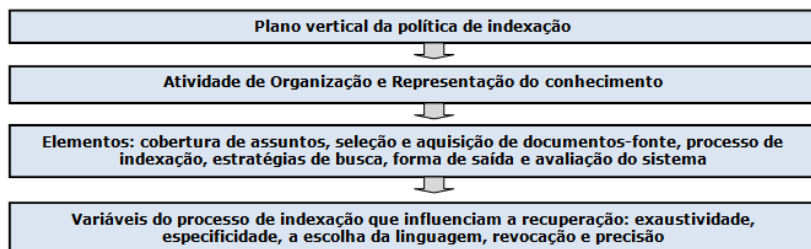


FIGURA 11 – Plano vertical da política de indexação.

Fonte: Baseado em Fujita (2012)

Com relação às diretrizes estabelecidas pela política de indexação, Fujita (2012, p. 22) argumenta que ganham significado quando aplicadas ao contexto de um sistema de armazenagem e recuperação da informação. A política torna-se uma importante aliada para que o profissional da informação realize o trabalho de representação temática dos documentos de maneira mais racional e consistente (RUBI, 2012).

Portanto, seguindo a linha de raciocínio de Strehl (1998), a definição de uma política de indexação deve levar em conta as características da organização em que está inserida e estar em plena consonância com os usuários a que se destina para que possa, assim, alcançar os objetivos de um sistema de informações: a recuperação das informações certas no momento adequado.

Mesmo com a implantação de diretrizes que venham a orientar a tomada de decisões com vista a facilitar a tarefa do indexador, alguns problemas são inerentes do indexador e independem (em menor ou maior grau) da política de indexação estabelecida.

No QUADRO 4 apresentam-se as vantagens e desvantagens entre o uso da indexação manual e da aplicação da indexação automática.

QUADRO 4 – Vantagens e desvantagens entre indexação manual e automática

	Indexação Manual	Indexação Automática
Vantagens	• A habilidade do indexador de contextualizar, relacionar palavras, usar a abstração, bem como decidir quais termos serão usados para identificar o conteúdo dos documentos para sua posterior recuperação.	• Maior rapidez no processo de indexação; • Pode complementar e aperfeiçoar o processo de indexação manual.
Desvantagens	• Processo de indexação mais lento; • Custo mais alto.	• Necessita do profissional da informação para a escolha das palavras-chave geradas pela mineração de textos.

Fonte: baseada em Araújo Júnior e Tarapanoff (2006)

Quanto a este assunto, Gil Leiva (2012, p. 82) cogita que na indexação automática não ocorrem fatores ligados ao indexador que podem afetar na decisão da escolha dos termos. Um programa de computador indexará sempre igual, bem ou mal, um documento sem que intervenha no contexto. A indexação variará somente quando forem feitas alterações nos parâmetros de análise do sistema. Portanto a Indexação Automática será o tema explorado na próxima seção.

2.3 Indexação Automática

Segundo Robredo (1982a, p. 238), “o processo de indexação automática se desenvolve seguindo um esquema bastante semelhante ao processo de leitura-memorização”, pelo qual o princípio geral da indexação automática está baseado na comparação de cada palavra do texto com uma relação de palavras vazias de significado pelo computador, previamente

estabelecidas, que conduz, por eliminação, a considerar as palavras restantes do texto como palavras significativas.

Deste modo, o histórico da indexação automática pode ser associado ao uso de programas computacionais para a geração de índices pré-coordenados (ROBREDO, 1982a).

De acordo com Borges (2009) a indexação automática visa à mecanização das atividades descritas anteriormente na seção 2.1, tendo como objetivo agilizar e auxiliar o processo intelectual realizado pelos profissionais da área.

Na visão de Vieira (1988a), o emprego da indexação automática é uma alternativa viável, face ao grande número de documentos existentes, tornando cada dia mais difícil realizar a indexação manual com um mínimo de qualidade requerida para assegurar o acesso posterior à informação.

Neste sentido a adição do computador para a tarefa de indexação segundo Gil Leiva (1997) procurou tornar mais ágil o processo da análise de informação, a obtenção de melhores índices de consistência e a redução dos custos, e, finalmente, uma maior qualidade nos sistemas de informação.

A indexação assistida por computador é considerada inicialmente um modelo de extração com características estatísticas e probabilísticas e sua origem coincide com as tentativas iniciais de junção da informática e da estatística com a área de documentação (BORGES, 2009).

A partir das ideias de Zipf, que em 1949, enunciou o princípio geral do mínimo esforço²⁴, que está relacionado com a tendência de repetição de certas palavras ao invés da utilização de palavras diferentes na comunicação oral ou escrita em uma língua (GIL LEIVA, 1997), Luhn sugeriu, em 1957, que a frequência de uma palavra em documentos está diretamente relacionada com a capacidade dessa palavra de/para representar o conteúdo do documento, no nível de indexação e de recuperação da informação.

Desta forma, Vieira (1988b), Cesarino e Pinto (1980) e Gomes (1989) observam que os primeiros anos de indexação automática baseada na frequência de palavras se desenvolveram na década de 1950, pois foi neste período que Luhn (1957) elaborou a idéia de que o vocabulário existente em um

²⁴ Segundo o qual a relação entre o produto da ordem de série (r) de uma palavra pela sua frequência de ocorrência (f) possui um valor constante (c) (GUEDES e BORSCHIVER, 2005).

documento deveria se constituir na base para a análise do seu conteúdo, sendo a melhor forma de recuperá-lo.

Esta análise ocorreria através da aplicação de um conceito básico e simples de indexação automática, o KWIC (Key Word in Context) ou palavra-chave no contexto baseado em permutações de palavras significativas nos títulos, resumos ou texto completo manipulado por máquina (PALMQUIST, 1998, tradução nossa).

Como explicam Vieira (1988b), Gil Leiva (1997) e Ferneda; Galvão e Rocha (2010), os primeiros métodos de indexação automática que surgiram nos anos 60 foram totalmente baseados em métodos estatísticos, mas posteriormente surgiram propostas de métodos híbridos que procuravam incorporar o conhecimento linguístico no processamento estatístico, pois se percebeu uma tendência de que os métodos híbridos possuem potencial para melhorar os resultados individuais conseguidos pelas abordagens linguísticas e estatísticas (CÂMARA JÚNIOR, 2013).

Na metade da década de 1960, Stevens (1965) apresentou em sua dissertação uma revisão dos critérios que levaram a aplicação dos computadores para a tarefa de indexação, e definiu a indexação automática como o uso de máquinas para extrair ou

atribuir termos de indexação sem a intervenção humana, uma vez que criaram programas ou regras sobre o procedimento.

Nos anos 70, a percepção da possibilidade de extrair os termos de um texto quando coincidem com algum tipo de dicionário armazenado de termos 'aceitáveis' foi a base de um importante trabalho sobre indexação com auxílio do computador realizado pelo Defense Documentation Center (LANCASTER, 2004).

Um método de ponderação, surgido também da década de 1970, consiste da discriminação do valor de termos, proposto por um grupo de investigadores liderados por Gerald Salton. Basicamente, a técnica classificava vocábulos de um texto segundo sua capacidade para diferenciar um documento de outro em uma dada coleção. De acordo com este método, são atribuídos pesos aos termos que, quanto mais altos os valores, significam que são os melhores termos para indexação (GIL LEIVA, 2012).

De acordo com Gil Leiva e Rodríguez Muñoz (1996a) até os anos 80 surgem técnicas não lingüísticas, baseadas na frequência das palavras, e se apoiando na probabilidade da relevância dos termos. Nesse contexto Narukawa e Fujita (2010) afirma que os primeiros sistemas de indexação automática foram baseados exclusivamente em métodos estatísticos e

probabilísticos, passando a incorporar métodos linguísticos somente a partir da década de 1980.

Nos anos 90, cresce a adoção do paradigma do Aprendizado de Máquina, um ramo da Inteligência Artificial, associado às técnicas de Processamento de Linguagem Natural (PLN). São os chamados métodos de indexação automática, baseados em frequência/ocorrência de palavras, que se constitui na construção de um classificador automático, capaz de inferir regras, de acordo com uma base de documentos já pré-classificada. Isto faz com que a Indexação Automática seja uma alternativa viável como aplicação para a Recuperação de Informação (GIL LEIVA, 1997).

Segundo Vieira (1988b) no final da 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, análise semântica e sintática do termo.

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 (GIL LEIVA, 1997).

Para 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 (KURAMOTO, 1995; SOUZA, 2006; BORGES; MACULAN; LIMA, 2008).

Segundo Lancaster (2004, p. 52), “vários programas de computador foram desenvolvidos para gerar automaticamente um conjunto de entradas de índice a partir de uma sequência de termos”. Alguns dos primeiros processos automatizados são os modelos KWIC, KWOC, KWAC, SLIC, PRECIS, POPSI, NEPHIS e o LIPHIS²⁵.

2.3.1 Fundamentos teóricos

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

²⁵ Um estudo aprofundado sobre os sistemas de indexação automática pode ser encontrado em Narukawa (2011).

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 uma 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 semi-automática – 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 destes conceitos como sinônimos.

Borges (2009, p. 31) menciona que a indexação automática “também chamada de indexação assistida por computador e de indexação semi-automática” 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 sentido 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 frequência de ocorrência. (NASCIMENTO, 2008, p. 24).

Nessa mesma linha, Hjørland (2008) define a indexação automática como a indexação realizada por meio de procedimentos algorítmicos.

A indexação automática de documentos é compreendida por Gomes (1989) e Santos e Ribeiro (2003) como o processo onde 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.

O embasamento da indexação automática consiste segundo Guinchat e Menou (1994, p. 182) "em fazer o computador reconhecer palavras que

aparecem no título, no resumo do documento, ou no seu próprio texto. Os termos reconhecidos são incorporados em um arquivo de pesquisa e servem para recuperar o documento.”.

O princípio geral da indexação automática baseia-se na comparação de cada palavra do texto com uma *stoplist*²⁶, previamente estabelecida, que conduz, por eliminação, a considerar as palavras restantes do texto como palavras significativas. O fluxograma da FIGURA 12 representa de maneira esquemática o processo básico de indexação automática de documentos (ROBREDO, 1982a).

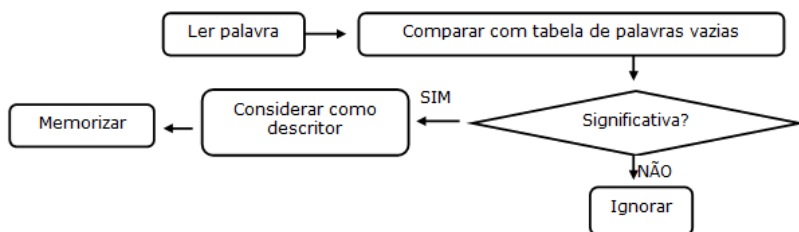


FIGURA 12 – Fluxograma simplificado do processo de indexação automática.

Fonte: Baseado em Robredo (1982a)

De acordo com Robredo (1982a), o processo de indexação automática pode identificar termos, pares de termos ou até frases significativas que expressem o conteúdo do documento. Na FIGURA 13 encontra-se um fluxograma mais detalhado comunicado por

²⁶ Relação de palavras vazias de significado (ROBREDO, 1982a).

Robredo (1982a) sobre o processo de Indexação Automática.

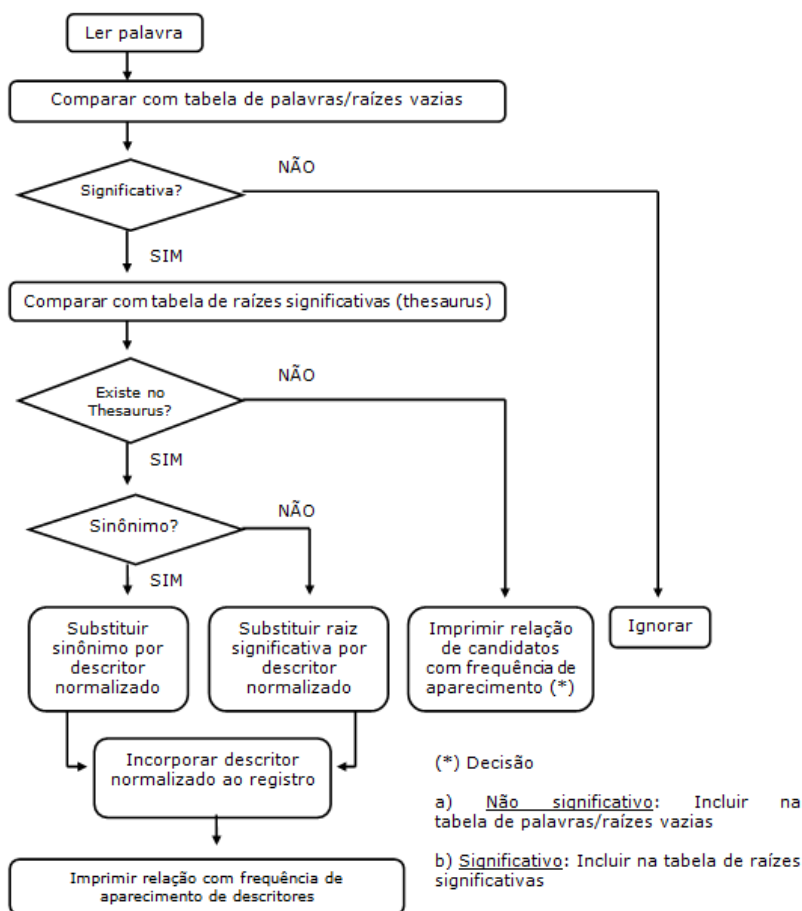


FIGURA 13 – Fluxograma mais detalhado do processo de indexação automática.

Fonte: Baseado em Robredo (1982a)

Tendo por base a aplicação dos computadores na representação automatizada dos documentos

Lancaster (2004) apresenta dois tipos de indexação automática:

1. A indexação automática por *extração* – ocorre quando o documento é indexado por um programa que adota critérios de frequência, posição e contexto com que as palavras aparecem no decorrer do texto para então extraí-las como descritores. Para obter os descritores, o programa obedece algumas regras como, por exemplo: eliminar os termos contidos na lista de *stopwords*, impossibilitando-os de se tornarem descritores. A desvantagem da utilização deste tipo de indexação está na possibilidade de obter descritores inadequados para representar o assunto abordado no documento.
2. A indexação automática por *atribuição* – consiste numa representação temática por meio de termos selecionados de um vocabulário controlado (tesauro ou lista alfabética). Mais complexa de ser executada pelo computador, por necessitar de mais sensibilidade em perceber e atribuir o termo mais adequado

como descritor. O modo de ser realizada por programas de computador é desenvolver, para cada termo a ser indexado um “perfil” de palavras ou expressões que costumam ocorrer frequentemente nos documentos. Esse tipo de perfil, por exemplo, para o termo chuva ácida, incluiria expressões como chuva ácida, precipitação ácida, poluição atmosférica, dióxido de enxofre etc. Mas segundo O’Connor (1965 apud LANCASTER, 2004), um dos problemas pode ser exemplificado na dificuldade do computador em atribuir o descritor, TOXICIDADE, ao texto: ‘dois dias depois de a substância haver sido ingerida surgiram diversos sintomas’, o que poderia facilmente ser percebida por um indexador humano.

O computador utiliza diferentes métodos para localizar os termos significativos através de uma abordagem realizada a partir de duas concepções diferentes, por vezes, complementares, os métodos não-linguísticos – que agrupam os estatísticos, a atribuição de pesos, os probabilísticos e os de análise de cluster – e, de outro lado, aqueles que decorrem de

certa análise lingüística dos textos (GIL LEIVA; RODRÍGUEZ MUÑOZ, 1996b).

Sobre este assunto Narukawa (2011, p. 56) comunica que “em contraposição à iniciativa da indexação automática, derivada exclusivamente de métodos estatísticos ou apenas de métodos lingüísticos, as propostas atuais são de integração de métodos”.

Na opinião de Vieira (1988b) as palavras significativas são selecionadas automaticamente, através de metodologias específicas, que são adotadas de acordo com as políticas de indexação e recuperação da informação, desenvolvimento *de software* e capacidade de *hardware* dos sistemas de informação.

Com base nestas considerações lembramos que a política de indexação automática deve estar inserida na política de indexação da biblioteca onde serão estabelecidos critérios para saber qual o processo/sistema/método melhor se enquadra (infraestrutura) para o perfil do usuário da biblioteca (destinatário) e dos documentos a serem indexados (contexto), pois em cada um dos setores de atuação de uma biblioteca devem existir políticas gerais e específicas funcionando como diretrizes administrativas de ação que orientem a tomada de decisão (RUBI, 2012).

Desta forma alguns métodos podem comparar os termos do resumo com os descritores da linguagem documental. Assim, cada vez que um descritor aparece no resumo, o documento é indexado com o termo correspondente (GUINCHAT; MENO, 1994).

Os autores ainda apresentam outros métodos, tais como: análises estatísticas em uma amostragem dos textos; determinação da frequência com que aparecem as palavras; os métodos lingüísticos que utilizam tratamentos morfológicos para o reconhecimento das estruturas mais significativas; e os métodos sintáticos que acrescentam ao método anterior uma interpretação sintática feita pelo computador, todos descritos no QUADRO 5.

QUADRO 5 – Métodos de Indexação Automática

Método	Descrição
Comparativo	• Compara através do computador os termos do resumo com os descritores da linguagem documental; • Cada vez que um descritor aparece no resumo, o documento é indexado com o termo correspondente. • Este método é eficaz apenas em campos muito específicos do conhecimento onde a linguagem natural e a linguagem documental são muito próximas.
Frequência ou análise estatística	• Realiza análises estatísticas em uma amostragem dos textos e determinam a frequência com que as palavras aparecem; • Neste caso as palavras são divididas em três grupos: em um se localizam as palavras irrelevantes, que aparecem constantemente; outro grupo é formado por palavras que aparecem com uma frequência menor do que a do grupo anterior e são consideradas como significativas; e o terceiro grupo formado pelas palavras que aparecem raramente e que são consideradas como muito específicas.
Lingüístico	• Introduz tratamentos morfológicos para o reconhecimento das estruturas mais significativas; • A indexação faz-se então em duas fases. A primeira consiste em determinar as unidades léxicas do documento a ser analisado. Em seguida, o computador faz o reconhecimento morfológico das palavras retidas (desinência verbal, grupo nominal), a partir de um arquivo de terminações e de ligações. As variantes morfológicas são reduzidas a uma forma léxica única. A segunda fase consiste em fazer o reconhecimento semântico dos termos retidos.
Sintático	• Acrescenta ao método lingüístico uma interpretação sintática feita pelo computador; • Os descritores isolados utilizados para a indexação são objeto de um tratamento gramatical que permite o estabelecimento de relações sintagmáticas entre os termos. Este método aproxima-se das operações efetuadas pelo homem, mas é mais difícil de ser executado.

Fonte: Adaptado de Guinchat e Menou (1994, p. 182-183)

Vieira (1988b) em uma revisão de literatura classificou os métodos de indexação automática como:

- 1) **Método de Frequência ou Análise estatística (Modelo Booleano)** – Este método trata do registro automático do aparecimento da palavra, que pode estar localizada no título, resumo, texto e em diversas combinações entre estas unidades, como, por exemplo, em título e resumo;
- 2) **Método de Atribuição de Peso (Modelo Vetorial)** – Este método atribui peso aos descritores como forma de atribuir-lhes valores semânticos para torná-los mais precisos, sem no entanto diminuir sua capacidade de revocação. É baseado na frequência de cada descritor;
- 3) **Método Probabilístico (Modelo Probabilístico)** – Este método aplica a frequência de co-ocorrência em palavras truncadas automaticamente. As palavras truncadas são extraídas através de um critério estatístico de distribuição binominal, denominado distribuição de Poisson. Aquelas

palavras cuja frequência de distribuição for descrita pela função de Poisson serão não-significantes.

- 4) **Análise de "Cluster"** – Este método se baseia no reconhecimento automático, por meio de procedimentos estatísticos que podem ser usados para classificar grupos de documentos por observação das semelhanças e dissemelhanças entre eles.
- 5) **Método de Associação entre Palavras ou Matemático** – Este método utiliza a frequência de ocorrência e co-ocorrência de palavras ou pares de palavras para identificar o conteúdo dos documentos. As palavras isoladas e as que se co-associam são identificadas em sentenças. Se as co-associações das mesmas palavras co-ocorrerem com determinada frequência, então, serão consideradas "descritores associados". Este método parte do princípio de que todas as palavras significativas estão relacionadas linearmente.

Para realizar a indexação automática de documentos é preciso que sejam realizadas operações

sobre o texto de cada documento. Estas operações tem a finalidade de se obter uma lista de termos que abranja todo o assunto abordado no documento e que façam parte das terminologias utilizadas pelos usuários que realizarão a consulta. Pouca utilidade há em um Sistema de Recuperação da Informação onde os termos apesar de abrangerem em sua extensão os assuntos do documento, sejam apresentados com termos pouco usuais pelos seus usuários.

Como parte do processo de indexar automaticamente, visando melhorar a recuperação do documento ocorre com a aplicação de operadores de texto sobre o conteúdo dos documentos.

Segundo Baeza-Yates e Ribeiro-Neto²⁷ (1999 apud GONZALEZ; LIMA, 2003), os principais operadores de texto (FIGURA 14) que permitem reduzir a quantidade da lista de termos, excluindo aqueles menos significativos, são: análise léxica – que elimina dígitos, pontuação etc.; eliminação de *stopwords* – artigos, preposições, pronomes, conjunções e alguns adjetivos e advérbios; operação de *stemming* – que reduz a palavra ao seu radical; seleção de termos de Indexação – determina as palavras que são utilizadas como elementos de

²⁷ Baeza-Yates, R.; Ribeiro-Neto, B. **Modern Information Retrieval**. New York: ACM Press, 1999. 513 p.

indexação (substantivos são mais representativos que adjetivos, verbos, advérbios, etc.). Estes operadores são empregados na indexação automática e podem ser utilizadas separadamente ou combinadas com a finalidade de melhorar os índices.

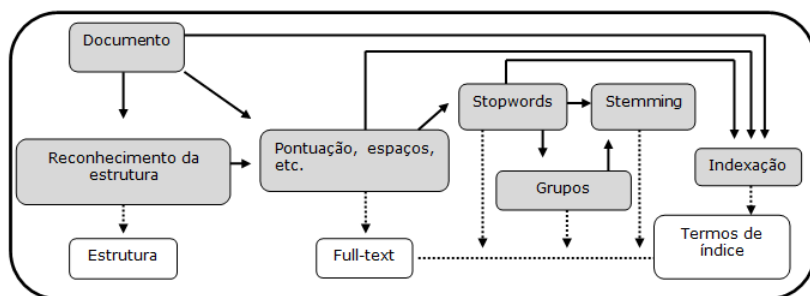


FIGURA 14 – Operações sobre o texto.

Fonte: Baseado em Baeza-Yates e Ribeiro-Neto (1999 *apud* GONZALEZ; LIMA, 2003)

2.3.2 Avaliação da indexação Automática

De acordo com Lancaster (1996), alguns autores destacam o papel da avaliação no processo decisório: a avaliação reúne dados necessários para determinar quais dentre várias estratégias alternativas parecem ter mais probabilidade de obter um resultado almejado. Outros consideram que a avaliação "é um componente importante da administração, pois é através dela que podemos corrigir ou manter estratégias com a finalidade de atingir objetivos pré-

determinados” (MACIEL; MENDONÇA, 2006, p. 23). Alguns afirmam que avaliação é um ramo da pesquisa – a aplicação do ‘método científico’ para determinar, por exemplo, a qualidade do desempenho de um programa. Este último conceito é o que nos interessa.

Lembra Lancaster (1996) que os critérios e processos empregados para avaliar uma busca por assunto numa base de dados em linha seriam igualmente aplicáveis à avaliação de uma busca por assunto num índice impresso e que com certas modificações também se aplicariam à avaliação de buscas por assuntos num catálogo em fichas (LANCASTER, 1996).

Neste sentido, os procedimentos que foram vistos na seção **2.1.3 Avaliação e qualidade na indexação** se aplicam também na Indexação Automática. Portanto, uma forma de avaliar a Indexação Automática é por meio dos cálculos dos coeficientes de precisão e revocação na recuperação de documentos de um sistema.

Outra fórmula é apresentada por Gil Leiva (1997), conhecida por coeficiente de avaliação e proposta inicialmente por Salton e McGill,²⁸ em 1983, para determinar a consistência entre a indexação

²⁸ SALTON, G.; MCGILL, M. J. **Introduction to modern information retrieval**. New York: McGraw-Hill, 1983.

automática e a manual. A fórmula com uma pequena modificação é a seguinte:

$$CI = \frac{T_{\infty}}{(A+B) - T_{\infty}}$$

Onde,

- CI = a coerência entre dois sistemas ou dois indexadores;
- T_{∞} = o número de termos comuns atribuídos pelos dois sistemas ou dois indexadores;
- A = o número de termos atribuídos pelo sistema 1 ou indexador 1;
- B = o número de termos propostos pelo sistema 2 ou indexador 2;

Esse coeficiente, explica Câmara Júnior (2007a), varia de valores entre "0" (zero) e "1" (um), onde o "0" é o pior caso, em que nenhum índice efetivamente coincidiu; e "1" o melhor caso, onde todos os índices atribuídos automaticamente foram idênticos aos índices manuais. Multiplicando o resultado por 100 tem-se a porcentagem de aderência entre a indexação manual e a automática.

Essa proposta foi inicialmente pensada para avaliar a consistência entre indexação manual e

indexação automática. Ela pode ser empregada para avaliação de sistemas de indexação automática, não no sentido de oposição entre a indexação manual e automática, mas no sentido de verificação da consistência para posterior melhora ou correção dos parâmetros do sistema de indexação, em um trabalho conjunto da indexação manual e automática (GIL LEIVA, 1997).

Uma das medidas utilizadas por Hlava (2002) para avaliar os sistemas de indexação automática refere-se à análise dos índices selecionados automaticamente. Essas medidas da precisão de um sistema baseiam-se em três categorias:

- Acertos – resultados que coincidem exatamente com o que o indexador humano teria aplicado ao sistema;
- Erros (misses) – palavras-chave que o indexador humano teria selecionado e, o sistema automatizado não o fez;
- Ruídos (noises) – palavras-chave selecionadas pelo computador que um indexador humano preferiria ter ignorado.

No estudo estatístico, dos valores dos acertos, erros e ruídos são a medida de avaliar a qualidade do

sistema. O percentual mínimo de acerto deve ser de 85% de um total de 100% de precisão de indexação humana para que o sistema seja considerado aceitável. Isso significa que a quantidade combinada dos índices, *Erros* e *Ruídos*, precisam ser inferiores a 15% (HLAVA, 2002).

Ainda de acordo com a autora, para que um sistema automatizado seja considerado bom precisará fornecer inicialmente, a partir de uma boa lista de palavras-chave, uma taxa de precisão de 60%, e com treinamento ou regras de construção um percentual de 85% ou mais. Isto significa que ainda há uma margem de erro esperado e que o sistema necessita – e melhora com – a avaliação humana.

Na opinião de Cesarino (1985) existe um número muito grande de linguagens de indexação sendo utilizadas em diferentes SRI. São variações das linguagens naturais x controladas; linguagens alfabéticas x classificatórias; linguagens pré x pós-coordenadas etc. Não se pode dizer, simplesmente, quais as melhores ou as piores. Não há linguagem perfeita. Existe sim, a linguagem que melhor se adapta a um SRI, com determinadas características.

Diante do que foi abordado, o referencial teórico a respeito da Indexação Automática se associa a esta

dissertação em seus aspectos conceituais como base para sustentar as análises propostas neste trabalho.

3 PROCEDIMENTOS METODOLÓGICOS

Neste capítulo, apresenta-se num primeiro momento as características metodológicas deste trabalho científico. Num segundo momento, apresenta-se o processo utilizado para a realização da seleção do *corpus* de análise. Na sequência, são apresentados os procedimentos prospectivos na realização do panorama sobre a indexação automática, bem como das análises bibliométricas e de conteúdo dos artigos.

3.1 Características metodológicas

Este estudo formou-se por meio do mapeamento e da 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 uma investigação dos resultados das análises bibliométricas, o que fornece ao estudo um caráter teórico.

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). Sendo assim, se propôs em realizar uma revisão de literatura com a finalidade de analisar os diversos aspectos referentes ao fato estudado. A produção científica nacional sobre a Indexação Automática é o que será analisado neste trabalho.

No que diz respeito aos procedimentos técnicos, se caracteriza como pesquisa bibliográfica, pois se trata do levantamento de toda bibliografia nacional já publicada, seja no formato de livros, revista, publicação avulsa e imprensa escrita (MARCONI; LAKATOS, 2010).

Esta dissertação se propõe a apresentar as ideias de pesquisadores, professores e cientistas brasileiros que atuam na área da Ciência da Informação sobre a Indexação Automática. Os métodos utilizados neste trabalho são Análise Bibliométrica e Análise de Conteúdo, como método de obter um panorama das pesquisas sobre o tema.

O método quantitativo utilizado para medir e avaliar o conhecimento científico do *corpus* desta dissertação é a Análise Bibliométrica, definida por Macias-Chapula (1998, p. 134) como “o estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada”.

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 uma 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).

A metodologia de análise de dados qualitativos utilizada nesta dissertação foi a Análise de Conteúdo, que constitui:

uma metodologia de pesquisa usada para descrever e interpretar o conteúdo de toda classe de documentos e textos. Essa análise, conduzindo a descrições sistemáticas, qualitativas ou quantitativas, ajuda a reinterpretar as mensagens e a atingir uma compreensão de seus significados num nível que vai além de uma leitura comum (MORAES, 1999, p. 7).

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 que os dados possam ser analisados e interpretados é preciso que estes sejam ordenados e

organizados. Para isto, devem ser codificados²⁹ e tabulados³⁰, começando-se o processo pela classificação³¹ (RUDIO, 1986). 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 da original, a fim de facilitar, num estado imediato, a sua consulta e referenciação.

3.2 Considerações sobre o corpus utilizado

Nesta seção, especifica-se a forma pela qual os documentos que compõem a revisão de literatura foram selecionados. Esclarece-se, também, como foi

²⁹ Codificar é o processo pelo qual se coloca uma determinada informação (ou, melhor, o “dado” que ela oferece) na categoria que lhe compete, atribuindo-se cada categoria a um item e dando-se, para cada item e para cada categoria, um símbolo (RUDIO, 1986, p. 124).

³⁰ Tabulação é o processo, pelo qual se apresentam graficamente os dados obtidos das categorias, permitindo sintetizar os dados de observação, de maneira a serem compreendidos e interpretados rapidamente e ensejando apreender-se com um só olhar as particularidades e relações dos mesmos (RUDIO, 1986).

³¹ Classificar é dividir um todo em partes (chamadas de categorias), sob um determinado critério, que é a base da divisão a ser feita, colocando cada uma das partes no seu lugar (RUDIO, 1986).

estabelecido o *corpus* para a concretização dos Objetivos Geral e Específicos.

Para a realização do referencial teórico e levantamento do *corpus* para as análises bibliométricas e de conteúdo referentes aos estudos sobre a Indexação Automática no âmbito da Ciência da Informação no Brasil foram utilizados documentos obtidos por meio de uma revisão de literatura através da coleta, seleção e leitura de livros, capítulos de livros, artigos de periódicos científicos, anais publicados em congressos e seminários, e literaturas cinzentas (monografias, teses e dissertações) localizadas em bibliotecas ou disponíveis na internet que abordem o tema do processo automático de representação da informação em documentos textuais. Para o referencial teórico observou-se documentos publicados nos idiomas português, inglês e espanhol, enquanto os 69 documentos que compõem o *corpus* de análise foram publicados no idioma português em âmbito brasileiro, e obtidos por meio de processo realizada por meio de pesquisa nas bases de dados virtuais BRAPCI³², Google Acadêmico³³ e PERI³⁴, e na biblioteca da UFPE, onde os documentos foram localizados através das expressões de busca

³² <http://www.brapci.ufpr.br>

³³ <http://scholar.google.com.br/schhp?hl=pt-BR&tab=ws>

³⁴ <http://bases.eci.ufmg.br/peri.htm>

'indexação automática', 'automatização da indexação' e 'indexação semi-automática'. O **Apêndice A** mostra o *corpus* que compõe o foco de análise para realização do panorama ordenado cronologicamente.

Os periódicos em que os artigos estão dispostos são:

- Transinformação;
- Ciência da Informação (Ci. Inf.);
- Informação & Sociedade: Estudos (Inf. & Soc.:Est.);
- AtoZ: Novas Práticas em Informação e Conhecimento (AtoZ);
- Revista de Biblioteconomia (R. Bibliotecon.);
- Biblionline;
- DataGramaZero;
- Perspectivas em Ciência da Informação (Perspect. Cienc. Inf.);
- Revista Escola de Biblioteconomia (R. Esc. Bibliotecon.);
- Revista Brasileira de Biblioteconomia e Documentação (RBBD);
- Estudos Avançados em Biblioteconomia e Ciência da Informação;

- Em Questão;
- Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação (Enc. Bibli: R. Eletr. Bibliotecon. Ci. Inf.).

Os anais foram comunicados nos seguintes congressos e seminários:

- Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB);
- Simpósio Latino-Americano de Terminologia;
- Encontro Nacional de Ciência da Informação (CINFORM);
- Seminário Nacional de Bibliotecas Universitárias (SNBU);
- Congresso Brasileiro de Biblioteconomia, Documentação e Ciência da Informação (CBBD).

E os livros publicados nas editoras:

- Briquet de Lemos / Livros;
- Thesaurus;
- Editora Universidade de Brasília;
- IBICT;
- Editora Ciência Moderna

O procedimento seguinte será a apreciação crítica das fontes de informação adquiridas

observando-as individualmente e em outro momento, separando-os em grupos de acordo com características comuns observadas através de indicadores pré-estabelecidos relacionados com as análises bibliométricas e de conteúdo apresentados nas próximas seções.

Uma medida adotada para facilitar as análises do *corpus* será a formatação de base de dados através do *software Microsoft Word*, em seguida utilizar-se-á o *software DataView*³⁵ 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.

3.3 Elaboração do panorama

Num 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

³⁵http://www.slb.com/services/characterization/software/data_utilities/dataview.aspx

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 apresentar o objetivo dos trabalhos, o 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.

Estes 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.

3.4 A metodologia prospectiva para a Análise Bibliométrica

Para realizar a análise bibliométrica o *corpus* a ser analisado será classificado dentro de sete categorias:

1. Autor – consta do nome do(s) autor(es) responsáveis pelo desenvolvimento intelectual da obra;
2. Instituição acadêmica – apresenta o nome da(s) Instituição(ões)/Organização(ões) envolvidas direta ou indiretamente na elaboração da obra;
3. Título – Título do trabalho;
4. Fonte – Tipo da fonte onde o trabalho foi publicado, se num livro, periódico, anais, tese, dissertação, capítulo de livro ou trabalho de conclusão de curso (TCC);
5. Instituição publicadora – Nome do Periódico/Evento/Editora onde o documento foi publicado;
6. Ano – Ano da publicação do trabalho;
7. Citação – Citação dos autores envolvidos na produção do trabalho.

Estas categorias serão representadas, respectivamente, pelas legendas: **AUT**, **INST**, **TIT**, **FONT**, **IPUB**, **ANO**, **CIT**. Sendo que as categorias TIT (título) e CIT (citação do trabalho) constarão somente como forma de orientação, não possuindo valor analítico, ver **APÊNDICE B**.

3.5 A metodologia prospectiva para a Análise de Conteúdo

No que se refere ao conteúdo, procura-se identificar para cada documento no *corpus* as seguintes categorias: o objetivo do trabalho e os aspectos metodológicos utilizados.

Na categoria destinada a analisar o objetivo, adotou-se o critério de que se o documento não propuser e nem aplicar um sistema/método/fórmula de indexação automática será automaticamente classificado como revisão de literatura. Desta forma os documentos podem ser classificados como:

- a) Revisão de literatura do assunto – os trabalhos marcados neste campo apresentam aspectos históricos e/ou os aspectos teóricos da indexação automática;

- b) Propõe – quando o trabalho propõe um determinado sistema/método/fórmula de indexação automática, portanto, neste campo encontram-se os trabalhos que realizam uma pesquisa sobre algum sistema/método/fórmula que pode ser aplicado na indexação automática, mas não há uma aplicação para comprovar sua eficácia;
- c) Aplica – o trabalho pretende comprovar a eficácia da aplicação de algum método/sistema/fórmula, neste campo estão localizados os trabalhos que além de pesquisar sobre a indexação automática se comprometem a comprovar a eficácia de sua aplicação.
- d) Aplica e propõe – quando o trabalho realiza a aplicação de algum método/sistema/fórmula com a finalidade de comprovar a eficácia de sua aplicação e propõe seu emprego.

Ainda na categoria objetivo, quando os trabalhos propõem ou aplicam, é classificado o nome do sistema/método/fórmula citado na pesquisa.

Em relação à segunda categoria, aspectos metodológicos, pretende-se analisar quais

procedimentos foram utilizados para avaliar os sistemas/métodos/fórmulas propostos pelos documentos através dos campos:

- a) Avaliação – se propõe a observar como ocorre o processo de avaliação da indexação automática, se através de comparação com a indexação manual, ou se houveram outras formas de comparação ou avaliação como forma de comprovar a eficácia do sistema/método/fórmula proposto;
- b) Natureza – identifica, de forma não excludente, se a análise ocorreu de um texto completo, resumo, citações, título ou verbetes;
- c) Tipologia – propõe apresentar o tipo de documento ao qual o sistema/método/fórmula se propõe a ser aplicado (carta, notícia, trabalho científico, atas etc.);

Além disso, na categoria dos aspectos metodológicos também encontram-se as análises dos sistemas de indexação automática, observando-se os campos:

- a) Validação dos termos – este campo busca analisar o tipo de indexação utilizada para

validar os termos, se assistida por computador, semi-automática, ou se tratava de uma indexação automática;

- b) Identificação dos termos – procura observar como ocorreu à identificação dos termos da indexação automática, se por extração ou por atribuição;
- c) Entrada de dados – tem a finalidade de observar qual o tipo de tratamento que o texto do documento analisado recebeu para a entrada de dados, se trata de:
 - Texto estruturado (em campos ou seções) – trata-se de uma linguagem de alto nível que utiliza campos para identificar entradas, ou valor de variáveis;
 - Um texto não-estruturado;
 - Um texto com marcas (marcando termos de indexação) – quando envolve a codificação simples de sequências de palavras em um arquivo de computador no formato texto-puro.

- d) Método/processo de identificação/ponderação/seleção de termos – busca verificar se a abordagem ocorreu por:
- Análise estatística – quando ocorre por 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.
 - Análise linguística – quando ocorre por análise morfológica, análise sintática, sintagmas nominais e/ou análise semântica.
 - Aprendizado por máquina³⁶ – quando ocorre por análise de cluster, e/ou classificação.
 - Representação do conhecimento³⁷ – quando ocorrer por gráficos conceituais, ontologia e/ou vocabulário controlado.

³⁶ Aprendizado de máquina ou "machine learning" é um termo empregado para representar quando os computadores são programados para aprender um determinado comportamento ou padrão automaticamente a partir de exemplos ou observações (MITCHELL, T. M. **Machine learning**. Mc Graw Hill, 1997).

- e) Linguagem de indexação – procura observar na:
- Natureza da linguagem – se foi utilizado uma linguagem natural (alfabética, pré-coordenada) ou uma linguagem controlada (classificatória, pós-coordenada).
 - Natureza dos termos – se os termos utilizados constam de palavras isoladas ou termos compostos.

Para responder as questões das categorias apresentadas procurou-se padronizar as respostas de forma a facilitar as análises posteriores, portanto foi estruturada uma ficha contendo os campos sugeridos para análise de conteúdo, que pode ser visualizada no **Apêndice C**.

Nesta ficha, as categorias/subcategorias e seus valores foram sendo identificados a partir da leitura do texto completo das publicações e da sistematização da descrição dos aspectos abordados em cada trabalho visando obter um panorama geral da produção científica sobre a temática.

³⁷ Termo empregado para representar os instrumentos utilizados na representação e organização do conhecimento (MOREIRA, A.; ALVARENGA, L.; OLIVEIRA, A. P. O nível do conhecimento e os instrumentos de representação: tesouros e ontologias. DataGramaZero - Revista de Ciência da Informação, v.5, n.6, dez. 2004.)

4 APRESENTAÇÃO DOS RESULTADOS E DISCUSSÃO

Esta seção estará dividida em duas subseções: na primeira, **4.1 Análises bibliométricas**, apresentam-se os resultados e as discussões relacionados às pesquisas bibliométricas realizadas no *corpus*. Para uma melhor compreensão dos estudos bibliométricos esta subseção encontrar-se-á subdividida em 4.1.1 Autores; 4.1.2 Instituições publicadoras; 4.1.3 Ano; 4.1.4 Fonte de publicação; e 4.1.5 Instituições acadêmicas; na subseção **4.2 Análise de conteúdo**, se encontram os resultados e as discussões relativas às suas categorias, subdivididas em 4.2.1 Objetivo; 4.2.2 Nome sistema/método/fórmula; 4.2.3 Avaliação; 4.2.4 Natureza do corpus; 4.2.5 Tipologia do corpus; 4.2.6 Validação dos termos; 4.2.7 Identificação dos termos; 4.2.8 Entrada de dados; 4.2.9 Linguagem de indexação; 4.2.10 Método/processo de identificação/ponderação/seleção de termos.

4.1 Análises bibliométricas

Nesta subseção estão localizados os resultados e as discussões a respeito das investigações

bibliométricas realizadas nos documentos que formam o *corpus* de pesquisa deste trabalho.

4.1.1 Autores

De maneira a observar a relação entre o autor e sua quantidade de pesquisas publicadas sobre a indexação automática, elaborou-se o GRÁFICO 1, que apresenta tanto a frequência de ocorrência de publicação de cada autor como a relação do seu hápax³⁸, utilizado nos resultados bibliométricos para denotar a quantidade de vezes que o termo foi utilizado só, sem a presença de outros termos nos resultados.

Analisando o GRÁFICO 1 podemos constatar que o resultado apresentado obedece a Lei de Lotka³⁹,

³⁸ É uma palavra ou expressão de origem grega que significa “uma só vez”, representando de que só existe uma única abonação nos registros da língua (ACADEMIA BRASILEIRA DE LETRAS, 2008).

³⁹ A Lei de Lotka está relacionada à produtividade dos autores e situasse na premissa básica elaborada por Voos em 1974 de que poucos pesquisadores publicam muito e muitos publicam pouco, seguindo a Lei do Inverso Quadrado $1/n^2$ onde em um dado período de tempo, analisando um número n de publicações, o número de pesquisadores que realizam duas publicações seria igual a $1/4$ do número de pesquisadores que escrevem um. O número de pesquisadores que fazem três publicações seria igual a $1/9$ do número de pesquisadores que escrevem um, e assim sucessivamente. (GUEDES e BORSCHIVER, 2005).

donde em um grupo de 78 pesquisadores levantados, apenas um pequeno número, formado por 17 pesquisadores, se destacam com duas ou mais publicações, desta forma averiguamos que os outros 61 pesquisadores restantes realizaram apenas uma única publicação. E daquele conjunto menor, apenas um seletor 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 SOUSA, 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.

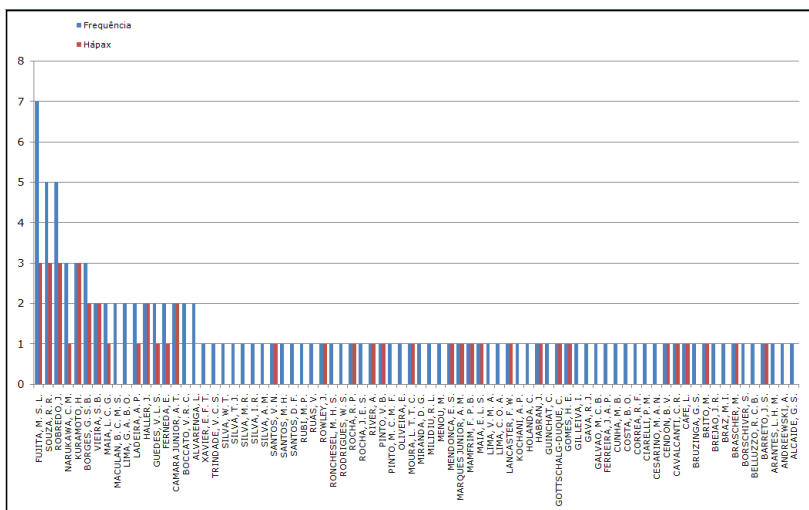


GRÁFICO 1 – Relação autor x quantidade de pesquisas publicadas
Fonte: desenvolvido pelo autor

Observando a produção deste reservado grupo de pesquisadores com três ou mais publicações constatamos que FUJITA, M. S. L.⁴⁰ apresentou um hápax de três, o que significa que das sete publicações realizadas por esta autora apenas três foram uma publicação individual, sendo, portanto, quatro de seus trabalhos produzidos com participação coletiva (com um ou mais autores).

Na sequência, tanto SOUZA, R. R.⁴¹ quanto ROBREDO, J.⁴², ambos com cinco publicações, apresentaram o mesmo valor de hápax, marcados na faixa dos três pontos, portanto cada um publicou três trabalhos individualmente, sendo dois realizados coletivamente.

O índice de hápax de NARUKAWA, C. M.⁴³ foi um, como sua produção consta de três trabalhos publicados, desta forma podemos inferir que dois foram desempenhados em consonância com outros autores enquanto um consta como publicação individual.

⁴⁰ <http://lattes.cnpq.br/6530346906709462>

⁴¹ <http://lattes.cnpq.br/4726949697973381>

⁴² <http://lattes.cnpq.br/9669125022187444>

⁴³ <http://lattes.cnpq.br/7858222747466392>

Em relação à KURAMOTO, H.⁴⁴ seu índice de hápax registra a marca de três de três publicações, o que neste caso significa que todas suas publicações foram realizadas individualmente.

BORGES, G. S. B.⁴⁵ com três trabalhos publicados e um índice de hápax dois, apresentando, portanto duas publicações individuais e uma coletiva é um caso diferente, pois se percebeu um erro causado pela utilização de dois sobrenomes distintos para a mesma pessoa, Graciane Silva Bruzuinga Borges, que também aparece em um trabalho como Graciane Silva Bruzuinga, acarretando desta forma que o mesmo autor seja interpretado como um autor diferente gerando índices que não corresponde a realidade prejudicando estudos que realizem análise de citação e estudos bibliométricos realizados sobre autores, desta forma, é importante que haja uma preocupação com a padronização dos nomes dos autores nos trabalhos.

Podemos ainda constatar no GRÁFICO 1, que no grupo formado por autores que realizaram apenas uma publicação apenas 20 deles apresentaram o hápax um, isto significa que apenas 20 de um total de 61 pesquisadores realizaram uma publicação individual e que os outros 41 autores com uma única publicação

⁴⁴ <http://lattes.cnpq.br/8778071496682261>

⁴⁵ <http://lattes.cnpq.br/1783747582273202>

a realizaram coletivamente. Esta constatação aponta para o baixo índice de continuidade por parte dos autores, onde apenas um pequeno grupo de 17 autores (aproximadamente 22%) realizou mais de uma pesquisa sobre o tema da indexação automática.

4.1.2 Instituições publicadoras

O GRÁFICO 2 foi preparado para permitir a visualização da relação entre os dados da frequência de publicação dos documentos do *corpus* com a instituição publicadora.

Nele, as publicações dos periódicos, editoras e anais descrevem uma curvatura decrescente até que ocorra uma padronização referente a uma única publicação.

As instituições publicadoras mais dedicadas ao assunto da automatização da indexação estão inseridas na área chamada de 'núcleo de publicação', onde encontram-se as informações consolidadas e que necessariamente quem é da área conhece.

Neste núcleo figura como destaque a revista 'Ciência da Informação'⁴⁶ com 17 trabalhos publicados sobre a temática, atualmente ainda está em vigor

⁴⁶ <http://revista.ibict.br/ciinf/index.php/ciinf>

publicando trabalhos inéditos relacionados com a ciência da informação ou que apresentem resultados de estudos e pesquisas sobre as atividades do setor de informação em ciência e tecnologia.

O segundo periódico com maior frequência é a Revista de Biblioteconomia de Brasília (RBB), que foi criada em 1973 e publicada até 2001, portanto no momento encontra-se fora de circulação, com sete artigos relacionados a indexação automática.

Na sequência estão os anais do ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO⁴⁷ (ENANCIB), fundada em 1989 e ainda em pleno funcionamento, constando de seis artigos publicados sobre o assunto.

A área da frequência de publicação 3 e 2 ilustram a região de mudança, onde os trabalhos começam a se afastar da concentração e é onde provavelmente localizam-se os meios de publicação que se constituem novidades e mudanças.

Os trabalhos localizados na área com frequência de publicação 1 pelo elevado índice de dispersão tem maiores chances de representar ruídos, contendo assim meios de publicação pouco consolidados.

⁴⁷ <http://www.ancib.org.br/pages/sobre.php>

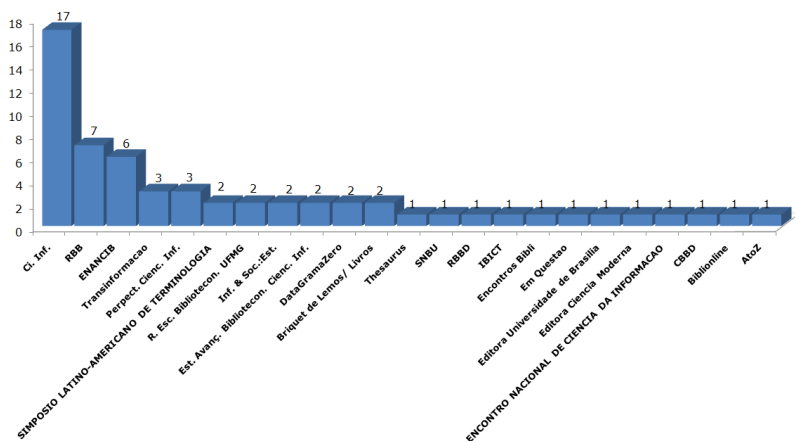


GRÁFICO 2 – Relação entre instituição publicadora e frequência de publicação

Fonte: desenvolvido pelo autor

Estabelecendo uma correlação entre as instituições publicadoras e os anos em que os trabalhos foram publicados, tendo por analogia a temporalidade restrita aos últimos 10 anos, obtemos como resultado a TABELA 1, de onde podemos observar que o ENANCIB foi a que mais publicou material referente à Indexação Automática nos últimos dez anos, com um total de seis trabalhos, distribuídos: um no ano 2005; dois em 2007; um em 2009; e outros dois em 2010.

TABELA 1 – Instituições publicadoras nos últimos dez anos

ANO	2004	2005	2006	2007	2008	2009	2010	2011	2012	Total
INSTITUIÇÃO PUBLICADORA										
Ciência da Informação	2			1						3
Transinformação	1									1
Em Questão	1									1
Briquet de Lemos / Livros	1									1
ENANCIB		1		2		1	2			6
CINFORM		1								1
Encontros Bibli			1							1
RBBD				1						1
Informação & Sociedade: Estudos					1	1				2
Perspectivas em Ciência da Informação						1	1			2
SNBU							1			1
CBBd								1		1
AtoZ								1		1
Ciência Moderna									1	1
Biblionline									1	1

Fonte: desenvolvido pelo autor

Em seguida, figura a revista Ciência da Informação, com três publicações, duas em 2004 e uma em 2007. Na terceira posição aparecem empatados, ambos com duas publicações, os periódicos: Informação & Sociedade: Estudos⁴⁸ (I&S), 2008 e 2009; e Perspectivas em Ciência da Informação⁴⁹, 2009 e 2010. As demais instituições publicadoras esboçam a marca de uma publicação cada.

⁴⁸ <http://www.ies.ufpb.br/ojs2/index.php/ies/index>

⁴⁹ <http://portaldeperiodicos.eci.ufmg.br/index.php/pci>

4.1.3 Ano de publicação

Dividindo a análise das publicações entre os séculos XX e o XXI constatamos que a diferença entre a quantidade de publicações é irrelevante, entre os anos 1970 e 1999 apresentaram-se a frequência de ocorrência de 29 publicações em contraste com as 40 publicações relacionadas aos anos 2000 a 2012. Entretanto, fazendo uma análise mais apurada, no século XXI, em apenas 12 anos (em comparação com os 29 relacionados ao século XX), a quantidade de publicação nacional sobre a indexação automática já excede em nove do número de publicações. Também observamos uma oscilação entre a frequência de ocorrência das publicações em relação ao ano, ilustrado no GRÁFICO 3, com alguns picos de cinco publicações nos anos 2004, 2007, 2009 e 2010.

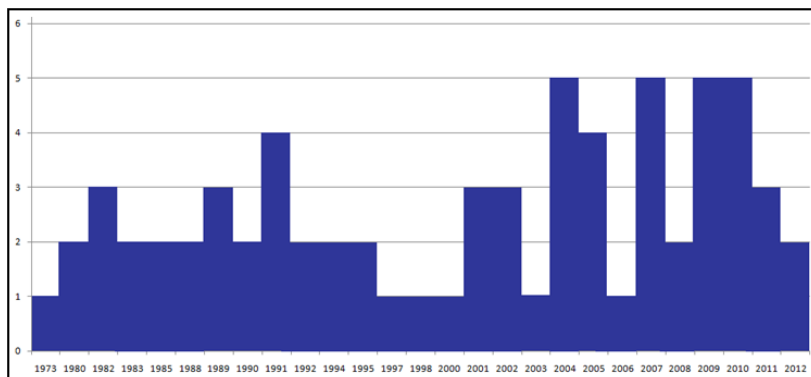


GRÁFICO 3 – Frequência de ocorrência dos trabalhos observados através do Ano de publicação

Fonte: desenvolvido pelo autor

Acreditamos que a tendência é aumentar o número de publicações nos próximos anos, pois de acordo com Estudillo García (2001) e Francelin (2004) 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.

4.1.4 Fonte de publicação

Com o objetivo de permitir a visualização das fontes de informação que mais realizam publicações sobre a indexação automática elaborou-se o GRÁFICO 4. Neste, podemos observar que os periódicos produzidos pelas revistas científicas da área de ciência da informação são as fontes publicadoras que tem maior destaque. Dos 69 documentos analisados 43 corresponderam aos periódicos, representando uma margem aproximada de 62%, 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.

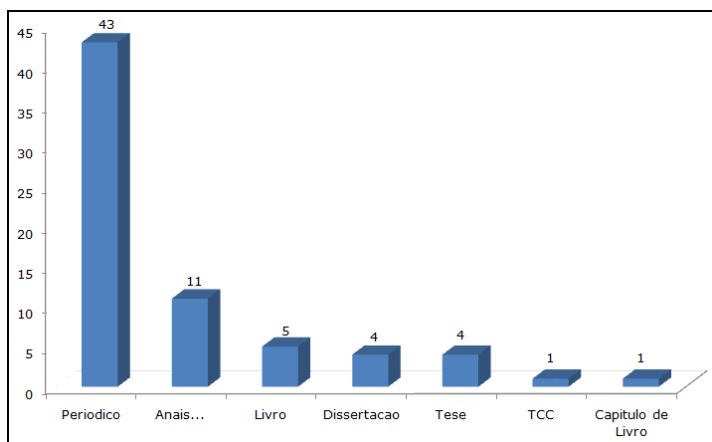


GRÁFICO 4 – Fonte de publicação
Fonte: desenvolvido pelo autor

Este fenômeno que possibilitou constataremos uma margem admirável na diferença entre a fonte de publicação que figura em primeiro lugar, o periódico, das demais fontes, pode ser justificado por meio de Cunha (2001) que atribuí às publicações periódicas ser um dos mais eficientes meios de registro e divulgação de pesquisas, estudos originais e outros tipos de trabalhos intelectuais. 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.

Os anais dos seminários, congressos e encontros da área de ciência da informação aparecem em segundo lugar com 11 documentos, representando

valores próximos dos 16%. A apresentação de trabalhos em eventos científicos (conferências, congressos, encontros, seminários, painéis, workshops, etc.) tem sido um dos mais importantes meios de disseminação de informação científica e tecnológica (ICT) (CUNHA, 2001).

Em seguida encontram-se os livros com 5 publicações o que corresponde a cerca de 7%. Uma fonte de informação secundária que “na área científica ou tecnológica normalmente serve para oferecer ao leitor um conjunto de conhecimentos consolidados sobre uma especialidade ou um estudo aprofundado de um tema restrito.” (CUNHA, 2001, p. 88).

Bem próximo estão as pesquisas publicadas em Dissertações e Teses, empatadas com quatro publicações, correspondendo a 6% cada. Nas palavras de Cunha (2001), as teses e dissertações são tipos de documentos que apresentam uma pesquisa original sobre determinado tema.

Por fim, foram relatados com uma única publicação, representando 1% cada, documentos em um trabalho de conclusão de curso e um capítulo de livro.

4.1.5 Instituições acadêmicas

O GRÁFICO 5 representa a frequência de ocorrência das instituições acadêmicas a qual os autores são vinculados.

Foram levantadas 28 instituições (nacionais e estrangeiras). Das seis instituições que apresentaram maior destaque, a Universidade de Brasília (UNB) está na primeira posição com uma frequência de ocorrência registrada em 16 documentos e o hápax registrando 13 documentos, logo 13 documentos foram produzidos apenas pela própria UNB, enquanto 3 documentos foram realizados em parceria com outras instituições: a Biblioteca Nacional de Agricultura (BINAGRI), a Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), e o Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (SEBRAE).

A instituição seguinte, a Universidade Federal de Minas Gerais (UFMG), apresenta valores idênticos de frequência de ocorrência e de hápax, 15, o significa que todos os quinze documentos produzidos pela UFMG foram realizados individualmente pela própria instituição sem a participação de outras.

Em seguida apresenta-se a Universidade Estadual Paulista (UNESP) com 10 documentos, destes, seis produções independentes e quatro com participação cooperativa com a Universidade de São

Paulo (USP), tendo em uma destas cooperações a participação da Universidade Federal de São Carlos (UFScar).

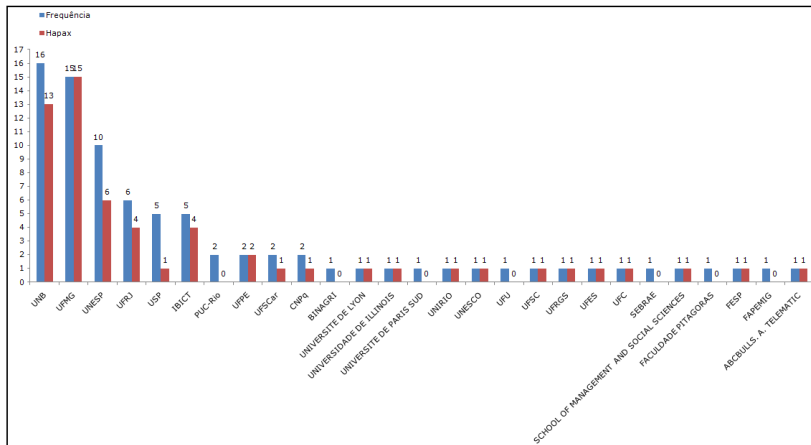


GRÁFICO 5 – Instituições acadêmicas
Fonte: desenvolvido pelo autor

Em quarta posição encontra-se a Universidade Federal do Rio de Janeiro (UFRJ) com seis pontos de frequência de ocorrência e quatro pontos marcando a produção individual. Dos dois trabalhos produzidos em parceria um foi com a Faculdade Pitágoras e o outro com o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Empatados com cinco documentos encontram-se a USP e o Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), diferenciando-se no número de trabalhos individuais, a primeira instituição

só realizou um trabalho sem participação de outras instituições e como apresentado anteriormente, houve parceria em todos os trabalhos com a UNESP, e uma vez com a UFSCar. Já o IBICT produziu quatro documentos individualmente e participou coletivamente de um trabalho com a instituição CNPq.

Observando a TABELA 2 podemos perceber que nos últimos dez anos, as três instituições que aparecem em primeiro lugar no GRÁFICO 5, ainda figuram como as maiores produtoras de documentos, agora a UFMG aparece com 11 trabalhos, a UNESP com 7, e a UnB 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.

TABELA 2 – Instituições acadêmicas nos últimos dez anos

ANO	2004	2005	2006	2007	2008	2009	2010	2011	2012	Total
INSTITUIÇÃO										
FAPEMIG	1									1
Universidade de Illinois	1									1
UNESP	1					2	2	1	1	7
UFU	1									1
UFRGS	1									1
UFMG	1	3		1	2	2	2			11
UFRJ		1	1				1			3
UNB				3						3
UFES				1						1
USP						2	2			4
UFSCar						1		1		2
Faculdade Pitágoras							1			1
UFPE								1	1	2

Fonte: desenvolvido pelo autor

A relação entre as instituições podem ser observadas no QUADRO 6 onde também se apresenta a frequência de relacionamento entre cada instituição. Através desta tabela observamos que a USP e a UNESP são as instituições que mais realizam pesquisas em co-participação.

QUADRO 6 - Relação entre as instituições acadêmicas

Relação das Instituições		Frequência
USP	UNESP	4
UFSCar	UNESP	1
IBICT	UFRJ	1
BINAGRI	UNB	1
USP	UFSCar	1
UNIVERSITE DE PARIS SUD	PUC-Rio	1
UNB	PUC-Rio	1
UFU	FAPEMIG	1
SEBRAE	UNB	1
Faculdade Pitágoras	UFRJ	1
CNPq	UFRJ	1
CNPq	IBICT	1

Fonte: desenvolvido pelo autor

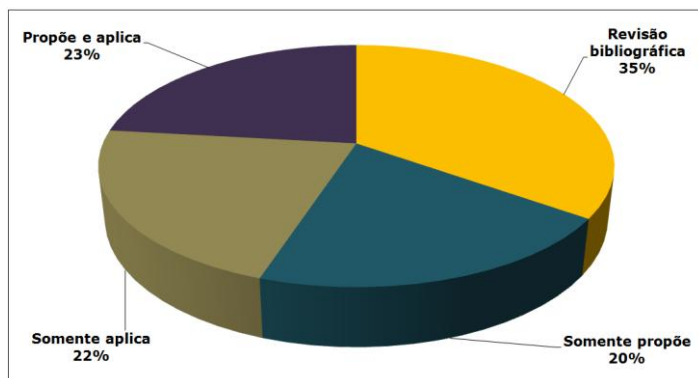
Verificou-se que as relações entre a USP e a UNESP ocorreram por fatores distintos, em uma das relações entre as instituições decorreu da relação docente e orientador (mestrado), em outra a relação foi motivada pelo doutoramento de docente, as outras relações provavelmente ocorreram devido a contatos interpessoais, motivados pelo estudo de temáticas semelhantes, realizados em eventos e ou em aulas/palestras em que foi convidado a ministrar.

4.2 Análise de conteúdo

Nesta subseção se encontram os resultados e as discussões relativas às categorias desenvolvidas para a análise de conteúdo, que estão localizadas nas subseções apresentadas a seguir.

4.2.1 Análise do objetivo dos documentos do corpus

Analisando o resultado do GRÁFICO 6 observamos que 24 trabalhos, o que corresponde a 35%, apresentam 'revisão bibliográfica' sobre o assunto da indexação automática ou algum tema relevante a recuperação de informação em que diretamente faz referencia à indexação automática.

GRÁFICO 6 – 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 20% do total, os outros 31 trabalhos estão divididos entre aqueles que 'aplicam', 15 trabalhos (21%), e os que 'aplicam e propõem', 16 trabalhos (23%), algum sistema, método ou fórmula.

Para cada trabalho analisado classificado como revisão bibliográfica, procurou-se identificar como abordam o tema da indexação automática.

Cesarino e Pinto (1980) apresentam a análise de assunto como operação-base para a recuperação da informação através de diferentes níveis: para definir o conteúdo de um documento; para compreender a necessidade de informação do usuário; para escolher ou criar a linguagem de indexação mais adequada para o sistema, além de apresentar a

possibilidade da análise de assunto automática de textos.

Cavalcanti (1982) apresenta os conceitos relacionados à indexação automática e alguns pesquisadores que se destacaram com trabalhos inovadores sobre esta temática, bem como classifica os diferentes tipos de indexação.

Vieira (1988b) analisa as pesquisas nacionais e estrangeiras que avaliam a qualidade da indexação manual e automática, em relação às técnicas e fontes empregadas para a extração dos termos significativos e a capacidade de recuperação da linguagem de indexação nas bases de dados.

Gomes (1989) realiza uma pesquisa bibliográfica sobre a indexação manual e a indexação automática e verifica que ainda não se trata de uma atividade tão automática, pois ainda depende da contribuição do profissional humano para melhorar o nível de recuperação.

Em Rivier (1992), apresentam-se os fundamentos teóricos das linguagens de indexação em estudos existentes na literatura especializada.

Guinchat e Menou (1994) realizam uma abordagem geral sobre as ciências e as técnicas da informação e documentação, abordando em um tópico

do capítulo 'A indexação' o assunto da indexação automática.

Em Mendonça (2000), apresenta-se a análise e a avaliação dos sistemas de indexação mediados pelo uso do computador e que se pautam na linguística como área de apoio à sua evolução, trazendo à pauta os assuntos da indexação automática e linguagem natural na área da automação e recuperação de documentos em três diferentes aspectos: as abordagens da indexação automática que estudam a aplicação operacional da estrutura de projetos; a recuperação automática e o processamento eletrônico da informação em base de dados; e o uso do Sintagma Nominal (SN) no encadeamento da informação.

Em Cendón (2001), oferece-se uma visão das principais categorias de ferramentas de busca na Internet, suas semelhanças, diferenças e características, fazendo uma relação entre a indexação automática e as primeiras tentativas de criar motores de busca por palavras-chave.

Pinto (2001) apresenta o estado da arte da indexação documentária, os seus fundamentos teóricos e a indexação manual.

Rowley (2002) apresenta alguns índices gerados por computador que são utilizados no processo da

indexação automática, tais como os índices *Kwic*, *Kwoc*, *Kwit* e *Kwac*, os índices rotados ou deslocados, *PRECIS* e o *Abstracts in New Technologies and Engineering* (ANTE), e o sistemas de indexação rotacional *Selective Listing in Combination* (SLIC).

Rocha (2004) analisa os mecanismos para organizar as informações da web, como classificadores automáticos, diretórios abertos, metadados, ontologias e Web Semântica.

Em Silva e Fujita (2004), identifica-se pela revisão de literatura, o desenvolvimento teórico e metodológico da prática do indexador e da evolução do processo de indexação.

Souza e Alvarenga (2004) identificam as tecnologias, as organizações associadas e o embasamento filosófico e conceitual subjacente a Web Semântica e suas contribuições na análise automática de documentos.

Em Lancaster (2004), dedica-se um capítulo específico para o tema da indexação automática, redação automática de resumos e processos afins, apresentando o processo histórico e definições da indexação automática.

Em Guedes e Borschiver (2005), desenvolve-se uma revisão da literatura sobre Bibliometria,

destacando sua aplicação como ferramenta estatística básica para a representação da informação e do conhecimento científico e tecnológico de forma automática.

Barreto (2007) apresenta os processos e métodos utilizados para a indexação e recuperação textual da informação semântica em vídeo, tendo como base a identificação e classificação do seu conteúdo visual e sonoro, onde sistemas híbridos, com indexação automática e análise de conteúdo supervisionada devem ser desenvolvidos.

Em Santos (2009), realiza-se um estudo sobre os modelos de indexação automática e seu uso no tratamento e recuperação de documentos na *Web* e em processos de indexação em bancos de dados bibliográficos.

Ladeira e Alvarenga (2009) e Ladeira (2010) analisam a produção científica nacional e contemporânea sobre o Processamento da Linguagem Natural, tendo em vista os desenvolvimentos realizados ao longo dos últimos 40 anos, o que inclui os tradutores automáticos, a indexação automática, a análise de conteúdo e estilo, além da recuperação de informação em documentos ou em textos

Em Borges (2009, 2010), desenvolvem-se os fundamentos históricos, teóricos e metodológicos da

indexação automática e realiza-se uma análise de critérios de indexação automática utilizados para o desenvolvimento de *softwares* destinados à extração de termos representativos do conteúdo de documentos textuais.

Kochani, Boccato e Rubi (2011) contribuem para o estabelecimento de uma política de indexação em sistemas automatizados que otimize o tratamento temático de *clippings*, pela perspectiva teórica e metodológica da área de Organização e Representação do Conhecimento.

Em Holanda e Braz (2012), identificam-se estratégias e procedimentos que os museus podem adotar ao disponibilizar conteúdos no ambiente Web para que sejam, posteriormente, recuperados de forma mais precisa quando buscados através do motor Google.

Ferneda (2012) avalia em quais aspectos a Ciência da Computação contribui para o avanço da Ciência da Informação, no que diz respeito ao processo de recuperação da informação, e associa a indexação automática à recuperação de informação no ambiente *Web*.

Levando em conta os trabalhos classificados como revisão bibliográfica, podemos constatar que: a maioria apresenta os fundamentos teóricos da

indexação automática, sua evolução histórica e desenvolvimento teórico metodológico; um segundo grupo se concentra em abordar a possibilidade de uso da automação da indexação na internet, por meio de motores de busca em estudos relacionados com a web semântica; alguns trabalhos comunicam 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 tendência teórica e prática para o aparecimento de outras pesquisas: o primeiro de 2005 que apresenta a aplicação das leis e princípios, com foco nas palavras, da bibliometria na indexação automática; um em 2007, que aborda a aplicação da indexação automática em vídeos; e outro em 2011 mencionando o desenvolvimento de uma política de indexação em sistemas automatizados.

Conforme podemos averiguar no GRÁFICO 7, 14 trabalhos foram classificados no grupo dos que propõem algum método, sistema ou fórmula. Destes, apenas um trabalho, correspondendo a 1%, realizou a proposição de fórmula, 11 trabalhos, o que corresponde a 37%, apresentaram a proposta de utilizar algum método, enquanto 18 (60%)

propuseram o uso de algum sistema de indexação automática.

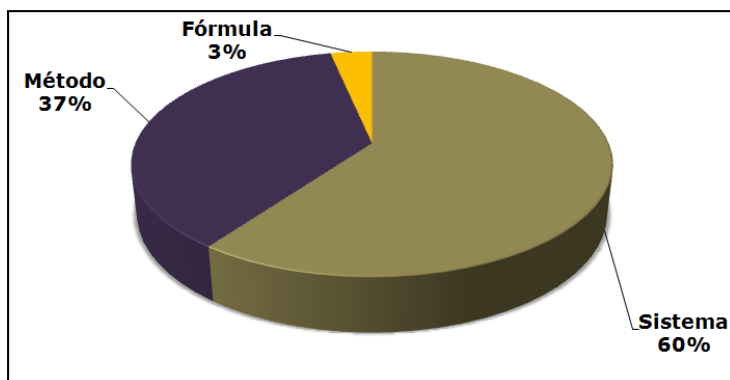


GRÁFICO 7 – Distribuição dos trabalhos que realizaram proposição
Fonte: desenvolvido pelo autor

Nos trabalhos rotulados como 'proposição', procurou-se congregar inicialmente aqueles que propuseram as fórmulas, em seguida os sistemas de indexação automática e na sequência os trabalhos que havia propostos métodos de indexação automática.

– Fórmula:

Em Mamfrim (1991) propõe-se a aplicação da indexação automática derivativa de textos em língua portuguesa, a partir de seu texto integral através da Fórmula de Transição de Goffman.

– Sistema:

Em Fujita (1989a), propõe-se o PRECIS como metodologia de indexação após avaliação da eficácia de recuperação do sistema.

Em Fujita (1989b), a proposição é do uso do sistema PRECIS na construção automática de índices de assunto.

O trabalho de Robredo e Ferreira (1980) e Robredo (1991) propõem a utilização do sistema de indexação automática, nomeado como AUTOMINDEX/II, para gerar uma tabela de frequências de aparecimento dos descritores encontrados em cada arquivo.

Andreewski e Ruas (1983) propõem o sistema de indexação automática SPIRIT, que aplica técnicas baseadas em métodos linguísticos e estatísticos, com o objetivo de processar os documentos em linguagem natural, permitindo assim o armazenamento e a interrogação em linguagem natural.

A proposta de Haller (1983, 1985) é a de um programa experimental de análise linguística, realizada de forma automática, de textos em linguagem natural em sistemas de informação.

Para Habran (1985) o sistema proposto, MISTRAL, permite realizar busca retrospectiva, disseminação seletiva da informação, entrada em linha

de documentos, indexação automática, gestão de léxico e tesouros, interface para videodisco, máquinas de microfilmagem e de fotocomposição.

Fujita (1990) propôs o PRECIS na compilação e seleção automática dos termos de indexação.

Em Belluzzo *et al.* (1990), a proposta é de realizar a indexação automática de documentos com o "PRECIS *software*".

Kuramoto (1997) propõe a construção de um Sistema de Recuperação de Informação Assistida por Computadores (SRIAC).

Gottschalg-Duque (2005) apresenta uma proposta para um Sistema de Recuperação de Informação baseado em Teorias da Linguística Computacional e Ontologia (SiRILiCO).

Em Maia (2008) e Maia e Souza (2010), propôs-se uma indexação automática dos documentos através dos sintagmas nominais.

A proposta de Narukawa e Fujita (2010) é a de investigar o uso e a adequação de Linguagens Documentárias alfabéticas em indexação automatizada na aplicação do *Sistema de Indización Semi-automática* (SISA).

Ferneda, Galvão e Rocha (2010) propõe um método de indexação automática de documentos através de um sistema computacional denominado *SintagMed*.

Narukawa (2011) propõe o Sistema de Indización Semi-automática (SISA) no processo de indexação.

Corrêa *et al.* (2011) apresenta como proposição o processo da indexação automática por meio de sintagmas nominais.

– Método:

Silva e Milidiú (1991) propõem um modelo de função de crença que possibilita indexação automática e o ordenamento em etapas dos documentos.

Em Kuramoto (1995), propõe-se uma abordagem alternativa de tratamento automático da linguagem natural (TALN) e acesso à informação na qual os SNs desempenham o principal papel.

Em Kuramoto (2002) o método proposto é o uso da organização hierárquica em árvore dos SNs na implementação dos termos de indexação automática para a recuperação de informação.

Em Café (2003), propôs-se como método para indexação automática investigar as Unidades

Terminológicas Complexas (UTC), que são fenômenos linguísticos que ocorrem em unidades léxicas que representam o assunto dos documentos.

Souza (2005a, 2005b, 2006) desenvolve uma proposta para indexação automática utilizando sintagmas nominais.

Em Câmara Júnior (2007a, 2007b), propõe-se a indexação automática por meio de Processamento de Linguagem Natural baseado nos métodos dos sintagmas nominais.

Em Bruzinga,^{*} Maculan e Lima (2007); Borges,^{*} Maculan e Lima (2008), apresentam como proposta de indexação automática o método de utilizar os SNs para obter os critérios sintático-semânticos.

Em virtude dos trabalhos classificados na categoria proposição, verificamos que a fórmula mais proposta foi o Ponto T de Goffman; em relação ao sistema, o mais citado foi o AUTOMINDEX/II; enquanto o uso dos Sintagmas Nominais foi o método mais proposto.

^{*} Os sobrenomes Bruzinga e Borges tratam da mesma pessoa, como explicado na análise bibliométrica desta dissertação na subseção 4.1.1 Autores, página 61.

Os trabalhos que são classificados com o objetivo de aplicar uma fórmula, sistema ou método na indexação automática, apresentaram as seguintes quantidades de trabalhos: três (10%), aplicam fórmula; 12 (39%), aplicam métodos e 16 (51%), aplicam sistema, ressaltados no GRÁFICO 8.

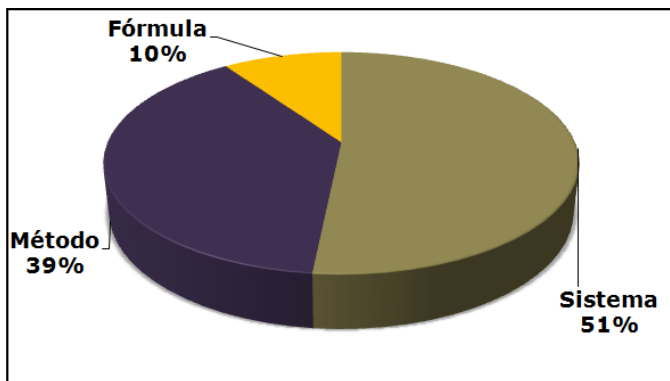


GRÁFICO 8 – Distribuição dos trabalhos que realizaram aplicação
Fonte: desenvolvido pelo autor

No grupo dos trabalhos que aplicam procurou-se inicialmente reuni-los de acordo com suas características em comuns, sendo apresentados inicialmente os que aplicaram fórmulas, em seguida os que aplicam sistemas, e na sequência aqueles onde se aplicou algum método na indexação automática.

– Fórmula:

Maia (1973) aplicou as leis de Zipf e a fórmula de transição de Goffman referentes às palavras de alta frequência de ocorrência em um texto, e observou que

possuem aplicações práticas na indexação automática de textos completos de documentos.

Em Mamfrim (1991), aplica-se a Fórmula de Transição de Goffman para a representação de conteúdo via indexação automática.

Em Guedes (1994), investiga-se a aplicação das fórmulas da primeira e da segunda lei de Zipf e a fórmula de transição de Goffman na indexação automática derivativa de textos científicos e tecnológicos.

– Sistema:

Em Robredo (1982b), aplicou-se o sistema AUTOMINDEX/II nas listas de controle de frequência de aparecimento dos descritores para aperfeiçoar os processos de indexação automática e recuperação de documentos.

Robredo (1982a) apresenta e aplica algumas funções do sistema AUTOMINDEX/II, na realização da indexação automática dos registros bibliográficos, a partir dos títulos ou resumos correspondentes, com ajuda de uma tabela de palavras vazias de significado e a de uma tabela de raízes significativas.

Vieira (1988a) aplica o sistema BIB/DIÁLOGO para realizar um estudo comparativo entre a qualidade da indexação manual e automática.

Fujita (1989a) aplica o sistema PRECIS para realizar uma avaliação da eficácia da recuperação dos índices gerados automaticamente pelo sistema, mediante estudo comparativo de uso dos catálogos de assunto da Biblioteca do Campus de Marília e do PRECIS.

Fujita (1989b) aplica o sistema de construção automático de índices de assunto em cadeia para investigar o funcionamento e adaptação do sistema PRECIS em língua portuguesa.

Em Fujita (1990), utilizou-se o sistema de indexação PRECIS para compilação e seleção automática dos termos de indexação.

Belluzzo *et al.* (1990) empregam o sistema PRECIS para realizar a indexação automática de documentos do acervo da Biblioteca da Faculdade de Odontologia de Bauru.

Marques Júnior (1991) apresenta a experiência de aplicar a indexação automática de atas de reuniões do Curso de Pós-Graduação em Biblioteconomia da UFMG.

Bräscher (2002) emprega o Zstation, um sistema de tratamento automático da linguagem natural, em um corpus constituído de documentos oficiais do Mercosul para testes de desambiguação.

Em Gottschalg-Duque (2005), construiu-se um protótipo de um Sistema de Recuperação de Informação baseado em Teorias da Lingüística Computacional e Ontologia, nomeado *SiRILiCO*, com a finalidade de apresentar análise e avaliação deste sistema de indexação automático e de linguagem natural, que é mediado pelo uso do computador e se pauta na área da lingüística como área de apoio.

Em Lima e Boccato (2009), utiliza-se o “Sistema de Indización Autmático” (SISA), para realizar a indexação automática e semi-automática, com o objetivo de comparar e analisar o desempenho terminológico dos processos de indexação manual, automática e semi-automática.

Narukawa, Gil Leiva e Fujita (2009) investigam a consistência na indexação manual e automática, e da exaustividade e precisão na recuperação da informação mediante aplicação do *software* SISA.

Em Ferneda, Galvão e Rocha (2010), implantou-se um sistema computacional desenvolvido a partir de um método estatístico de sintagmas para indexação automática, denominado *SintagMed*.

Maia (2008) e Maia e Souza (2010) utilizaram o OGMA para automatizar a extração dos SNs e o cálculo do peso de cada termo na indexação dos documentos para cada um dos seis métodos propostos.

Em Narukawa (2011), aplica-se *ThesAgro* no *Sistema de Indización Semi-Automática* (SISA) para realizar análise comparativa com a indexação manual da Biblioteca Nacional de Agricultura (BINAGRI).

Em Corrêa *et al.* (2011), utiliza-se o OGMA para analisar o texto e analisar a aplicabilidade da extração de SNs na indexação automática e recuperação de informação.

– Método:

Brito (1992) aplica o método de descrever o conteúdo informacional dos textos através dos SNs no tratamento automático de linguagens naturais para melhorar sua indexação.

Moura (1995) emprega o método de frequência de ocorrência de palavras para extração e ordenação das palavras significativas para realizar uma análise da indexação automática.

Em Robredo e Cunha (1998), mostrou-se o interesse em estudar os métodos de análise da co-ocorrência de palavras ou expressões significativas para 'mapear' um determinado campo do conhecimento.

Alcaide *et al.* (2001) destacam que utilizam uma representação automática baseada em um método estatístico-morfológico, onde a extração dos léxicos é

feita com o auxílio de um dicionário que possui somente uma lista de significados e termos compostos.

Em Silva, A., Silva, I. e Arantes (2004), implanta-se o método de frequência de ocorrência das palavras encontradas nos textos para a indexação automática de todas as palavras para que possa gerar no momento da consulta um ranking em ordem decrescente de similaridade da consulta.

Em Souza (2005a, 2005b, 2006), aplica-se um método para viabilizar o processo de atribuição automática de descritores a textos digitalizados utilizando SNs.

Oliveira *et al.* (2007) enfatizam a representação dos documentos baseados no método de frequência de ocorrência no tratamento automático de textos.

Câmara Júnior (2007a, 2007b), aplica um modelo de indexação automática de acórdãos baseado nos métodos dos SNs.

Levando em conta a aplicação, as fórmulas bibliométricas citadas foram as leis de Zipf e o ponto T de Goffman, em relação aos sistemas os mais comentados foram o AUTOMINDEX/II e o SISA, já o método baseado na frequência de ocorrência foi o que obteve maior menção.

O GRÁFICO 9 ilustra os trabalhos que realizam aplicações e proposições de fórmula, sistema ou método no processo automático de indexar termos, onde um (6%) trabalho aplica e destaca a viabilidade do uso de uma fórmula matemática; cinco (31%) trabalhos estão relacionados com os métodos; e 10 (63%) trabalhos aplicam e propõem sistemas de indexação automática.

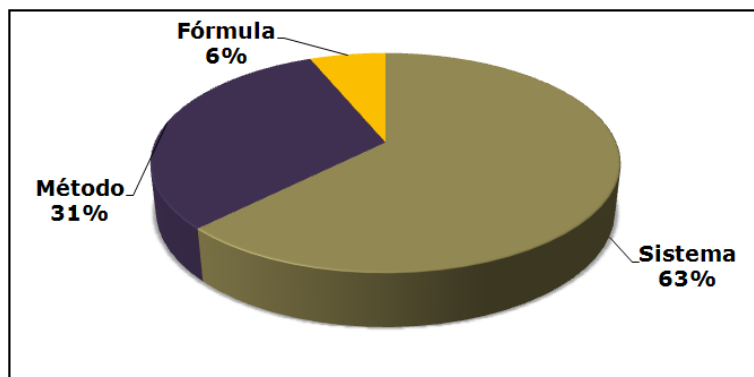


GRÁFICO 9 – Distribuição dos trabalhos que realizaram proposição e aplicação

Fonte: desenvolvido pelo autor

Em seguida procurou-se agrupar os trabalhos a partir de suas peculiaridades e demonstrar quais foram a fórmula, sistema ou método propostos pelos autores.

– Fórmula:

Em Mamfrim (1991) aplica-se e propõe-se a Fórmula de Transição de Goffman na indexação

automática derivativa de textos integrais em língua portuguesa.

– Sistema:

Fujita (1989a) aplica e propõe o sistema PRECIS na recuperação dos índices gerados automaticamente.

Em Fujita (1989b), a aplicação e proposição do sistema automático de índices de assunto é por meio do sistema PRECIS.

Em Fujita (1990), utilizou-se e propôs-se para compilação e seleção automática dos termos de indexação o sistema de indexação PRECIS.

Belluzzo *et al.* (1990) empregam e propõem o sistema PRECIS na realização da indexação automática.

Gottschalg-Duque (2005) aplica e propõe um protótipo de um Sistema de Recuperação de Informação baseado em Teorias da Lingüística Computacional e Ontologia, o *SiRILiCO*.

Em Fernalda, Galvão e Rocha (2010), implantou-se e propôs-se o *SintagMed*, um sistema computacional desenvolvido a partir de um método estatístico de sintagmas para indexação automática.

Maia (2008) e Maia e Souza (2010) utilizaram e propuseram o OGMA para automatizar a extração dos sintagmas nominais.

Em Narukawa (2011), aplica-se e propõe-se o *ThesAgro* no *Sistema de Indización Semi-Automática* (SISA).

Corrêa *et al.* (2011), usa e propõe o OGMA na extração de sintagmas nominais para a indexação automática e recuperação de informação.

– Método:

Em Souza (2005a, 2005b, 2006), aplica-se e propõe-se a utilização dos sintagmas nominais como método para viabilizar o processo de atribuição automática de descritores.

Câmara Júnior (2007a, 2007b), aplica e propõe um modelo de indexação automática de acórdãos baseado nos métodos dos sintagmas nominais.

Em razão dos trabalhos que aplicam e propõem constatou-se um único trabalho referente à fórmula de transição de Goffman, no que tange aos sistemas o PRECIS e o OGMA foram os mais citados, enquanto aos métodos, todos fizeram menção aos Sintagmas Nominais.

4.2.2 Nome Sistema/Método/Fórmula

As fórmulas localizadas nas pesquisas são as Leis de Zipf, observadas em dois trabalhos, e o Ponto T de Goffman, que aparece como assunto em três, ambas consiste em fórmulas bibliométricas relacionadas com a frequência de ocorrência de palavras em textos.

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 10, de onde averiguamos que os três sistemas mais pesquisados para representar automaticamente os descritores aparecem empatados com quatro (17%) trabalhos cada um, são eles: o sistema SISA, o PRECIS e o AUTOMINDEX/II. Logo em seguida aparece o OGMA, sendo pesquisado por três trabalhos, correspondendo a aproximadamente 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 CPD da 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 sendo 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 aplicado e proposto por apenas um trabalho. Desta forma os sistemas SintagMed, Zstation, SRIAC, SPIRIT, MISTRAL e SiRILiCo, representando 60% dos sistemas observados, não apresentam continuidade nas pesquisas e foram concebidos 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 prosseguimento nas pesquisas sobre estes sistemas.

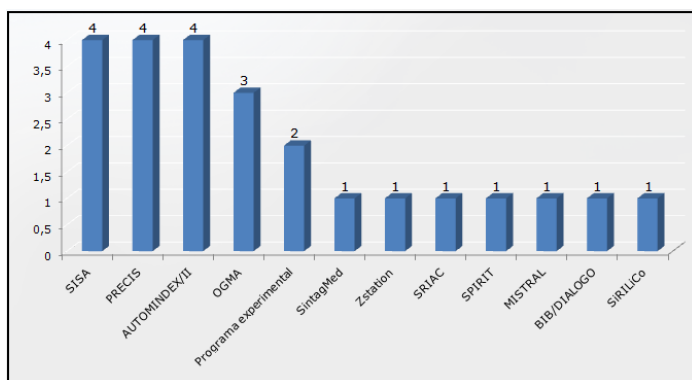


GRÁFICO 10 – Frequência de ocorrência dos sistemas nos documentos do *corpus*

Fonte: desenvolvido pelo autor

Avaliando os resultados referentes ao GRÁFICO 11 obtemos que o método mais pesquisado para representar automaticamente os descritores foi através da aplicação da frequência de ocorrência dos termos, com 24 (53%) trabalhos, dividido entre 10 trabalhos que aplicaram; sete que propuseram ou aplicaram; e sete que propuseram um método. Em seguida encontra-se o SN com 17 (38%) trabalhos, organizados em dois trabalhos que aplicam, cinco que propõem, e 10 que aplicam e propõem. Os demais métodos apareceram apenas uma vez, foram eles:

função de crença, frequência de co-ocorrência, estatístico-morfológico e UTC.

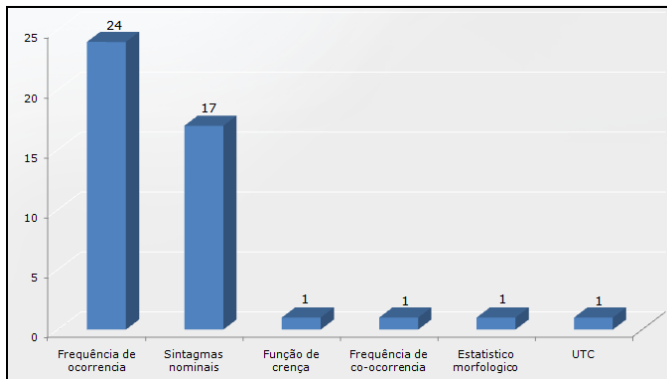


GRÁFICO 11 – Frequência de ocorrência dos métodos nos documentos do *corpus*

Fonte: desenvolvido pelo autor

Fazendo uma relação dos nomes dos Sistemas/Métodos/Fórmulas com o ano em que os trabalhos foram publicados conseguimos observar os dez últimos anos, que não constam nos resultados os trabalhos que aplicaram e/ou aplicaram e propuseram fórmulas bibliométricas revelando que se trata de trabalhos antigos (o último datado de 1994, vinte anos atrás).

No recorte temporal dos sistemas, o SISA aparece como o mais bem colocado, com quatro (44%) dos trabalhos, o OGMA com três (33%), e empatados, os sistemas SintagMed e SiRILico, cada

um com um (11%) trabalho, ilustrado pelo GRÁFICO 12.

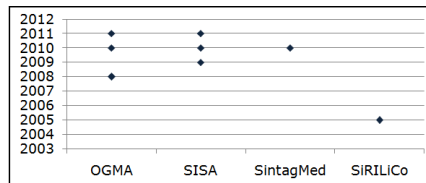


GRÁFICO 12 – Relação do Nome do Sistema observados nos últimos dez anos

Fonte: desenvolvido pelo autor

Os sistemas SISA e OGMA merecem uma atenção maior de trabalhos futuros, pois são os que figuram entre os sistemas mais pesquisados nos últimos dez anos.

O GRÁFICO 13 registra o estudo em relação aos métodos pesquisados nos últimos dez anos revelando que os trabalhos se dividiram em duas linhas de estudos, em primeiro lugar os SN com 12 trabalhos e em segundo com dez trabalhos a frequência de ocorrência.

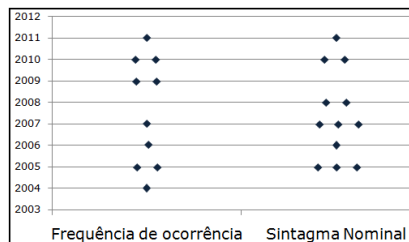


GRÁFICO 13 – Relação do Nome do Método observados nos últimos dez anos

Fonte: desenvolvido pelo autor

4.2.3 Distribuição da avaliação dos documentos

De acordo com os dados contidos no GRÁFICO 14, 24 (53%) trabalhos avaliaram sistemas, 18 (40%) métodos e três (7%) fórmulas.

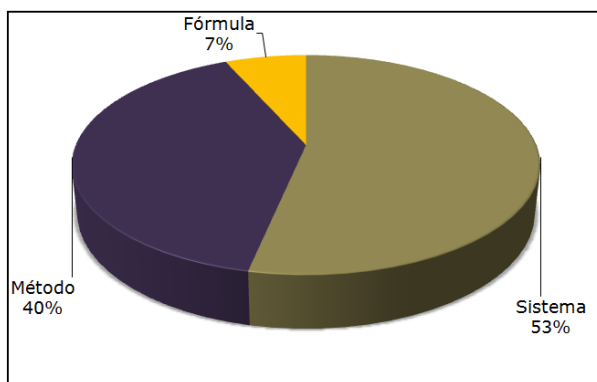


GRÁFICO 14 – Distribuição da avaliação dos documentos
Fonte: desenvolvido pelo autor

Para a grande maioria dos autores o método escolhido para avaliar suas pesquisas sobre o processo automático da indexação e assuntos afins foi realizando uma comparação dos resultados obtidos pelo processo automático com os obtidos através do método intelectual de indexação, correspondendo a 27 (71%) dos trabalhos, conforme GRÁFICO 15. Consideramos que a maioria dos trabalhos optou como método de avaliação a comparação com a indexação manual, por ser o método mais comum utilizado.

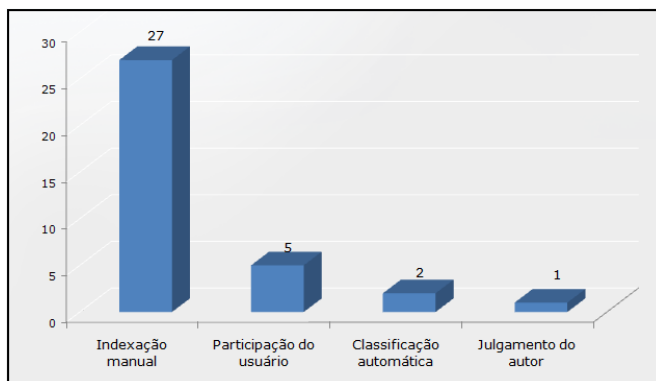


GRÁFICO 15 – Frequência de ocorrência dos métodos de avaliação
Fonte: desenvolvido pelo autor

Em segundo lugar, encontra-se o método de avaliar através da participação do usuário na comparação de índices ou SRIs, com seis (14%) trabalhos. Neste processo geralmente se aplica, nos 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.

Com dois (11%) trabalhos, encontra-se 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.

4.2.4 Natureza do corpus

Quanto à natureza do *corpus*, constatou-se que ocorreu uma preferência em se realizar pesquisas com o texto completo dos documentos, em 29 (69%) trabalhos que tratam da recuperação da informação através da indexação automática.

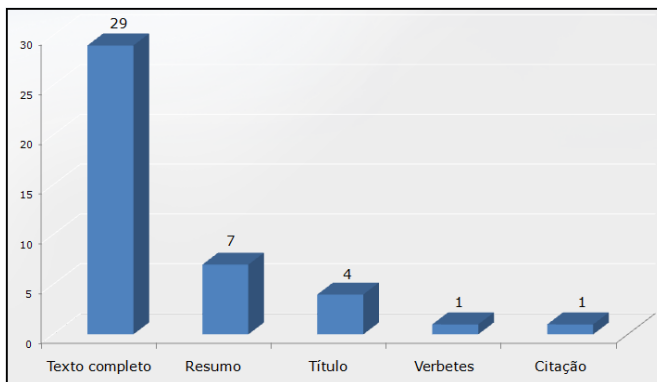


GRÁFICO 16 – Natureza do corpus

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 realizaram

em suas pesquisas como *corpus* de análise os títulos, e tanto os verbetes quanto as citações constam como *corpus* de investigação em apenas um trabalho, cada. Os resultados estão ressaltados visualmente no GRÁFICO 16.

4.2.5 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 17 podemos constatar que poucas tipologias ocorreram muitas vezes, enquanto diversas tipologias ocorreram poucas vezes.

Os dados ilustrados no GRÁFICO 17 ajudam a compreender qual foi o comportamento da natureza do *corpus* pesquisado nos trabalhos analisados nesta Dissertação. Verifica-se que a tipologia mais pesquisada com 27 trabalhos, correspondendo a uma margem de 66%, foram os trabalhos científicos, pois julgamos ser o tipo de material que mais interessa as instituições, que normalmente desenvolvem e/ou financiam estas pesquisas, como visto na subseção 4.1.5 Instituições, tratam-se de Universidades Públicas em sua grande maioria, e por este motivo a

importância da natureza do *corpus* incide sobre os trabalhos científicos que normalmente são produzidos na própria instituição.

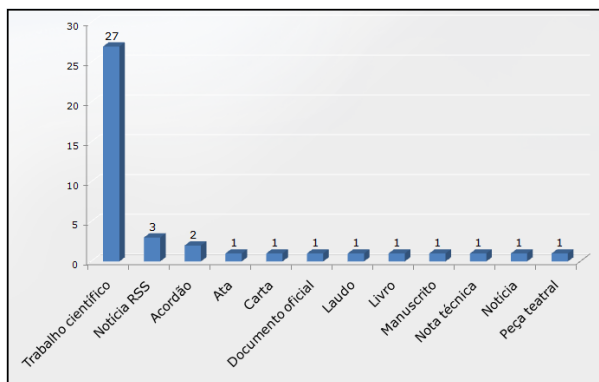


GRÁFICO 17 – Tipologia do corpus

Fonte: desenvolvido pelo autor

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

4.2.6 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 18 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.

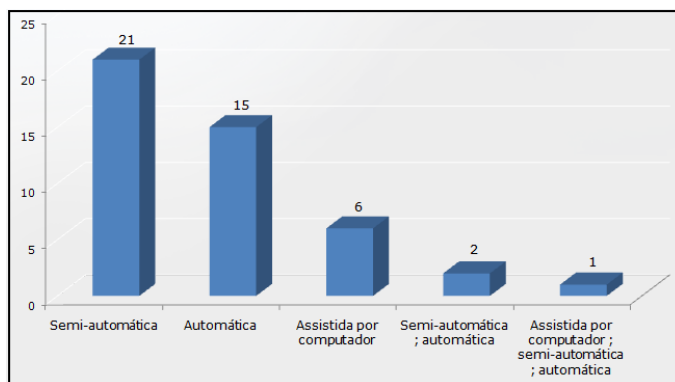


GRÁFICO 18 – Validação dos termos

Fonte: desenvolvido pelo autor

Realizando o QUADRO 7 contendo a relação do Ano com o Tipo de validação do termo, dos últimos dez anos, podemos constatar que os trabalhos que validam os termos utilizando a indexação automática passaram a superar aqueles que o fazem através do processo semi-automático, entretanto esta diferença não é muito significativa. Segundo os dados apresentados na tabela, o método de indexação

automática fez parte de onze (52%) trabalhos, o semi-automático foi contabilizado em oito (38%) estudos, e dois (10%) citam a indexação assistida por computador.

QUADRO 7 – Relação Ano e Tipo de validação do termo dos últimos dez anos

Ano	Frequência		
	Assistida	Semi-automática	Automática
2004	1	-	-
2005	-	2	1
2006	-	1	-
2007	-	1	4
2008	-	-	2
2009	1	2	1
2010	-	1	2
2011	-	1	1
TOTAL	2	8	11

Fonte: desenvolvido pelo autor

4.2.7 Identificação dos termos

O GRÁFICO 19 demonstra que a diferença entre os trabalhos que realizaram estudos do processo de identificação dos termos possui valores bem próximos, com 22 (49%) trabalhos realizando o método por meio da extração, enquanto 23 (53%) adotaram o processo através da atribuição.

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.

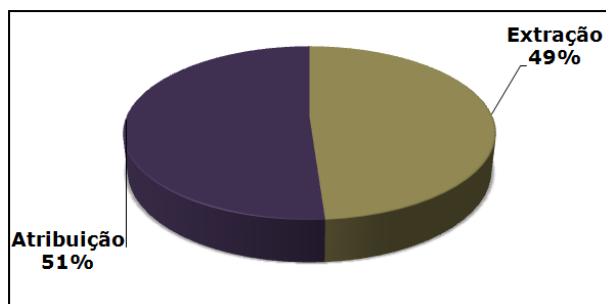


GRÁFICO 19 – Identificação dos termos
Fonte: desenvolvido pelo autor

4.2.8 Forma que ocorreu a entrada dos dados

O GRÁFICO 20 ilustra como o ocorreu à entrada dos dados analisados pelos trabalhos que compõem o *corpus* desta dissertação. 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.

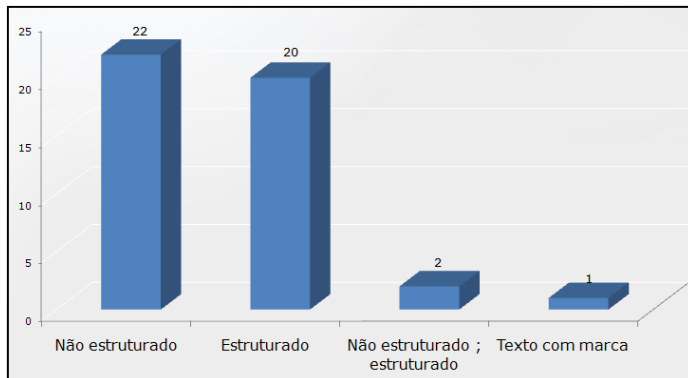


GRÁFICO 20 – Frequência de ocorrência da forma como ocorreu a entrada de dados

Fonte: desenvolvido pelo autor

4.2.9 Linguagem de indexação

Nesta subseção procurou-se decompor os trabalhos quanto a sua natureza em linguagem (natural ou controlada) e em termos (palavras isoladas ou termos compostos).

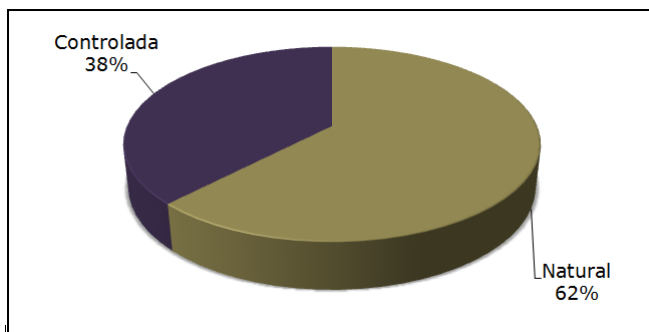


GRÁFICO 21 – Natureza da linguagem

Fonte: desenvolvido pelo autor

Quanto à natureza da linguagem notamos que ocorreu uma predominância pela linguagem natural com 28 (62%) trabalhos em decorrência dos 17 (38%) trabalhos atribuídos à linguagem controlada, visualizados no GRÁFICO 21.

Os dados relacionados aos termos estão ilustrados no GRÁFICO 22, que ressalta a superioridade numérica dos termos compostos com 29 (64%) trabalhos em relação à escolha apenas por palavras isoladas com 16 (36%) trabalhos.

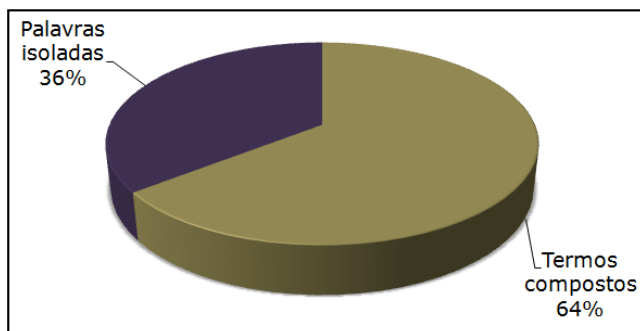


GRÁFICO 22 – Natureza dos termos

Fonte: desenvolvido pelo autor

4.2.10 Método/processo de identificação / ponderação / seleção de termos

Os métodos ou os processos utilizados para identificar, ponderar ou selecionar os termos utilizados na indexação automática (em um total de 45

documentos que propuseram, aplicaram ou aplicaram e propuseram analisados) foram observados quanto ao uso da análise estatística, análise linguística, aprendizado por máquina e representação do conhecimento, sendo classificados em categorias e mensurados quanto à frequência em que apareceram nos trabalhos, estas categorias estão descritas na seção **3.5 A metodologia prospectiva para a Análise de Conteúdo**. O resultado pode ser observado no GRÁFICO 23, onde o método mais pesquisado foi a análise estatística com 34 (35%) trabalhos, o segundo método com maior pontuação foi a análise linguística, aparecendo em 29 (30%) dos trabalhos, a representação do conhecimento está em 25 (26%) trabalhos 27%, enquanto o aprendizado de máquina condiz com oito trabalhos, reproduzido aproximadamente em 9%.

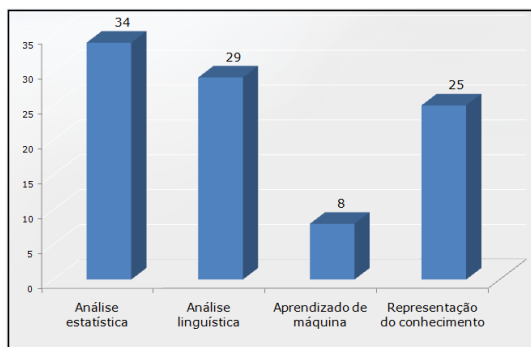


GRÁFICO 23 – Frequência de ocorrência do método/processo de identificação/ponderação/seleção de termos
Fonte: desenvolvido pelo autor

De acordo com a ilustração do GRÁFICO 24, podemos ratificar que a categoria mais utilizada foi a frequência de ocorrência, em 30 trabalhos, correspondendo a aproximadamente 15%, o vocabulário controlado aparece em 24 (12%) trabalhos, empatados na terceira posição estão as stopwords e a análise semântica com 23 (12%, aproximadamente) trabalhos. Logo em seguida, citada por 22 trabalhos, correspondendo a 11% das categorias, figura a análise sintática.

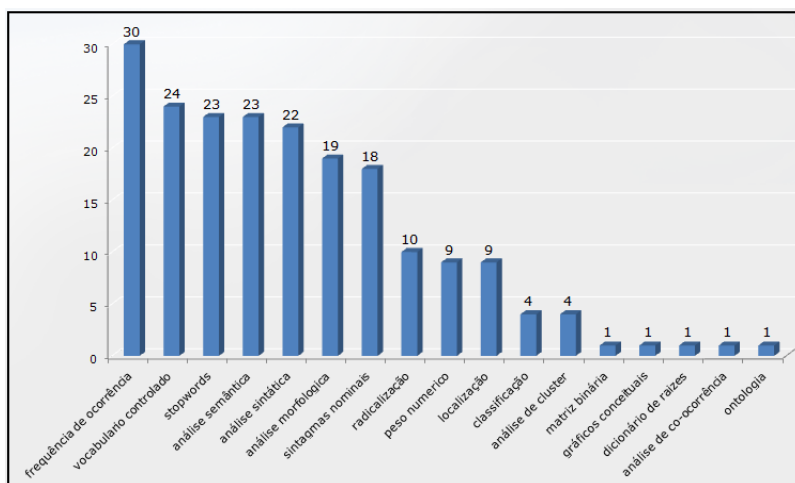


GRÁFICO 24 – Procedimentos internos tomados pelo sistema
Fonte: desenvolvido pelo autor

Ainda visualizando o GRÁFICO 24 observa-se que a análise morfológica é apontada por 19 (10%) trabalhos, enquanto o sintagma nominal é assinalado em 18 (9%) trabalhos. Na sequência ainda constam a

radicalização com 10 (5%), o peso numérico e a localização, cada um com nove (aproximadamente, 5%) trabalhos, a classificação e a análise de cluster com quatro (2%) trabalhos, cada, e nas categorias relatadas uma única vez estão a matriz binária, os gráficos conceituais, o dicionário de raízes, a análise de co-ocorrência e a ontologia.

Ao analisar a TABELA 3, desenvolvida através da relação dos dados obtidos através da frequência das categorias utilizadas na análise de conteúdo do *corpus* com o ano dos documentos, podemos averiguar que:

- No início da década de 70 as pesquisas em Ciência da Informação (CI) estão orientadas para a estatística apresentando interesse nas pesquisas referentes à frequência de palavras particularmente com relação à Bibliometria;
- Algumas categorias, já consolidadas no tratamento da indexação automática, como a frequência de ocorrência e o vocabulário controlado são recorrentes nas pesquisas sobre este tema a partir de

1980 até os dias atuais. Pode-se apontar como causa o fato dos pesquisadores e estudiosos procurarem buscar realizar melhorias nas técnicas e nos instrumentos utilizados no processo da indexação automática permitindo que ao longo dos anos sofram mudanças para se ajustarem às novas exigências proporcionadas pelo surgimento de novas tecnologias;

- As publicações na área da ciência da informação sobre a aplicação do Processamento da Linguagem Natural (PLN) na indexação automática tem início a partir da década de 80 com o auxílio da combinação dos métodos linguísticos com os métodos estatísticos, mas é no início dos anos 2000, que há uma retomada nas publicações sobre o PLN, associado às pesquisas sobre a aplicação dos Sintagmas Nominais (SN) como método de indexação automática, possibilitado pelo desenvolvimento tecnológico que

permitiu o surgimento de novos softwares e de computadores mais potentes tornando viável a análise linguística de um texto completo.

TABELA 3 – Apreciação das categorias utilizadas na análise de conteúdo do corpus em relação ao ano

Método		Análise estatística							Análise linguística				Aprendizado de máquina		Representação do conhecimento			
Categorias		Radicalização	Frequência de ocorrência	Stopwords	Localização	Frequência de co-ocorrência	Peso numérico	Dicionário de raízes	Matriz Binária	Análise Morfológica	Análise Semântica	Análise Sintática	Sintagmas Nominais	Análise de cluster	Classificação	Gráficos conceituais	Ontologia	Vocabulário controlado
Documento	Ano																	
1	1973		x		x													
2	1980	x	x	x	x													x
4	1982	x	x	x	x													x
5	1982	x	x	x	x													x
7	1983	x	x	x			x			x	x	x			x			x
8	1983	x	x	x			x			x	x	x			x			x
9	1985	x	x	x						x	x	x						x
10	1985	x	x	x			x			x	x	x			x			x
12	1988	x	x	x	x													
14	1989		x								x	x						x
67	1989		x								x	x						x
15	1990		x								x	x						x
16	1990		x								x	x						x
17	1991		x		x													
18	1991																	x
19	1991	x	x	x	x	x				x	x	x						
20	1991		x				x				x							x
21	1992									x	x	x	x					
23	1994		x		x													
24	1995	x	x	x	x													
25	1995												x		x			

26	1997												x					
27	1998		x	x		x								x				
29	2001		x	x						x								x
32	2002												x					
33	2002									x	x	x	x				x	x
68	2003									x	x	x	x					
35	2004		x															
38	2005		x	x						x	x	x	x					
40	2005		x	x						x	x	x	x					
69	2005									x	x	x	x				x	
41	2006		x	x						x	x	x	x					
42	2007									x	x	x	x					x
43	2007						x			x	x	x	x	x				x
44	2007		x	x			x											
46	2007						x			x	x	x	x					x
47	2008									x	x	x	x					x
48	2008			x			x			x	x	x	x	x				
50	2009		x	x														x
51	2009		x	x														x
54	2010			x			x			x	x	x	x	x				
55	2010		x	x					x				x					x
56	2010		x	x														x
59	2011												x					
60	2011		x	x														x

Fonte: desenvolvido pelo autor

5 CONCLUSÃO

Em um primeiro momento, após a leitura dos documentos que compõem o *corpus* desta dissertação, verificamos que muitos pesquisadores realizam estudos sobre a Indexação Automática como solução para melhorar o processo de recuperação da informação. Isso ocorre devido à motivação deles em suprir os efeitos dos fenômenos proporcionados tanto pelo crescimento vertiginoso da produção de documentos - que aumenta em progressão geométrica-, quanto pelas consequências geradas pela subjetividade, fruto da indexação realizada por humanos.

Baseando-se nesses dados, este trabalho visa descrever e analisar a produção científica desses estudiosos, professores e cientistas no que tange à indexação automática no Brasil, entre os anos 1973 a 2012, delimitados à vista da Ciência da Informação (CI). Para tanto, ele procura identificar quem é autoridade no assunto, qual periódico é referência na publicação de artigos sobre o tema, bem como o período apontado com a maior distribuição temporal das publicações. Objetiva também saber qual fonte de informação é considerada pelos autores como o principal meio de divulgação, a instituição acadêmica

que apresenta maior produção, além de distinguir quais foram os objetivos, os aspectos metodológicos e os sistemas, métodos e fórmulas presentes no *corpus*.

A partir dessa problemática, elaborou-se um panorama sobre a temática por meio de duas abordagens. A primeira é qualitativa, através da seleção de categorias para serem analisadas na produção acadêmica e científica sobre a Indexação Automática no campo da Ciência da Informação. Já a segunda, ocorre por uma abordagem quantitativa, proveniente de uma análise dos resultados das análises bibliométricas.

Como justificativa para a realização deste trabalho, citamos o valor da informação obtida da análise do *corpus* levantado, permitindo identificar as tendências e realizar projeções sobre futuras pesquisas.

Os resultados indicam que na área de Ciência da Informação, nos estudos quanto ao assunto da indexação automática, há certa concentração de publicações por parte de autores, instituição publicadora, escolha por parte dos autores do tipo de fonte de informação para divulgar seus trabalhos e instituições acadêmicas. Por esse motivo, muitos dos resultados vislumbrados ao longo deste trabalho obedeceram à distribuição de palavras em um texto

conforme o padrão das distribuições bibliométricas em geral: “poucos com muito e muitos com pouco”.

Verificamos também que a autora com maior número de publicações sobre a indexação automática foi Fujita, M. S. L., com sete trabalhos, dentre os quais três apresentaram autoria única. Além disso, constatamos um erro no valor atribuído a uma das autoras proporcionado pelo uso, em trabalhos distintos, de codinomes diferentes para a mesma pessoa. É imprescindível que ocorra uma padronização dos nomes dos autores para que os resultados não sejam deturpados.

Ainda em relação à produtividade dos autores nos veículos analisados, concluímos que a maior parte da produção acadêmica se referiu a autores que publicaram apenas um artigo, representando 78% da produção acadêmica analisada nesta dissertação, o que denota indícios de falta de continuidade nas pesquisas.

Por fim, constatamos que a revista *Ciência da Informação* (Ci. Inf.), uma publicação quadrimestral de trabalhos inéditos relacionados com a ciência da informação, destacou-se como o periódico que mais realizou publicações sobre a automatização na indexação, com 17 artigos, dez a mais do que o segundo colocado.

Com base nos resultados apresentados em relação ao estudo bibliométrico sobre a distribuição temporal das publicações, verificamos a tendência de um aumento nos estudos e, conseqüentemente, nas publicações do processo automático nos sistemas de recuperação da informação. Estes são alavancados através de estudos originais baseados no surgimento de novas tecnologias da informação e da comunicação e do aprimoramento das que já existem.

Constatamos que a principal fonte de informação utilizada pelos autores para divulgação de seus trabalhos acerca da indexação automática foram as publicações periódicas produzidas nas revistas científicas. Isso se deve ao fato de elas serem um tipo de publicação primária, fácil de ser acessada pela comunidade científica, o que a torna importante nas áreas de ciência e tecnologia.

Em relação às instituições, averiguamos que a Universidade de Brasília (UNB) foi a instituição que mais produziu trabalhos relacionados com o processo automático da indexação com um total de 16 documentos, sendo três em co-participação com as instituições Biblioteca Nacional de Agricultura (BINAGRI), Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), e Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (SEBRAE).

Por outro lado, nos últimos anos percebemos que novas instituições começam a demonstrar interesse em pesquisar a seleção automática de descritores na área da Ciência da Informação, tais como a USP, a UFRJ, a UFPE e a UFScar, entre outras.

Fundamentados nos resultados das análises de conteúdo do *corpus* desta dissertação, observamos que as categorias analisadas nos objetivos apresentaram percentuais próximos. Além disso, percebemos que a maior parte dos trabalhos classificados como revisão bibliográfica realiza uma abordagem da história da indexação automática, apontando seus fundamentos teóricos e evolução dos métodos.

De acordo com os resultados, podemos concluir que os sintagmas nominais e a frequência de ocorrência foram os métodos mais citados. Em relação aos sistemas de indexação automática, dois merecem uma atenção especial, o SISA e o OGMA, pois são os mais pesquisados desde os anos 2004. Esses dados ratificam a importância de estudar os referidos métodos e sistemas com a finalidade de aplicá-los na extração automática de descritores como forma de melhorar a recuperação de informação.

Verificamos que a grande concentração dos trabalhos utilizou como método de avaliação a

comparação com a indexação manual por ser o método mais comum atualmente, apesar dos problemas que são levantados pelos autores. Dessa forma, eles procuram avaliar se a implantação do sistema automático trará benefícios, mesmo se os valores da avaliação forem similares, através do ganho de tempo.

Baseados nos dados obtidos através da análise de conteúdo quanto à natureza e a tipologia do *corpus*, avaliamos que a preocupação da maioria dos autores está concentrada em indexar o texto completo de trabalhos científicos. Esse fato possibilita uma contribuição significativa para o desenvolvimento da sociedade ao propiciar a reflexão crítica e as buscas por paradigmas sociais que inspiram as organizações e os indivíduos.

Analisando o tipo de indexação que os trabalhos estudados nesta dissertação realizaram para validar os termos de suas pesquisas, 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 (47%) e os que optaram pelo da atribuição (53%), há uma explicação provável. É que a chegada de novas tecnologias e de pesquisas sobre aplicação de tesouros e vocabulários controlados motivaram 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, com a diferença apenas de dois trabalho a favor dos textos não estruturados em relação ao texto estruturado. Em relação à linguagem de indexação, foi observada tanto a natureza da linguagem, que demonstrou uma preferência dos trabalhos pela pesquisa com a linguagem natural, quanto a natureza dos termos, no qual a primazia encontra-se no estudo com termos compostos.

Quanto à categoria dos métodos e processos de identificação, ponderação e seleção dos termos, chegamos à seguinte constatação. Tanto o vocabulário controlado, categoria referente à representação do conhecimento, quanto a frequência de ocorrência, categoria referente à análise estatística, são

recorrentes nas pesquisas sobre a indexação automática desde a década de 80 até os dias atuais. Uma das razões pode ser devido aos pesquisadores buscarem realizar melhorias nas técnicas e nos instrumentos que já vêm sendo aplicados.

Este trabalho não possui um ponto de chegada, ou término, pois o movimento ininterrupto da ciência continuará incentivando os pesquisadores (tanto os mais recentes quanto os antigos) 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 superior a realizada pelo especialista humano quando realizam a mesma tarefa.

Em vista disso, este trabalho aponta para trabalhos futuros na área da Ciência da Informação, sobre a indexação automática, que possam se beneficiar das contribuições apresentadas e possibilitem, dessa forma, dar continuidade ao trabalho desenvolvido nesta dissertação.

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;
- Traçar o panorama da produção científica 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

ACADEMIA BRASILEIRA DE LETRAS. **Dicionário escolar da língua portuguesa**. 2. ed. São Paulo: Companhia Editora Nacional, 2008.

ALCAIDE, G. S. et al. Análise comparativa e de consistência entre representações automática e manual de informações documentárias, **Transinformação**, v. 13, n. 1, p. 23-41, jan./jun. 2001.

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.

ANÍZIO, J. L. A.; NASCIMENTO, G. F. C. L. Avaliação do processo de indexação na biblioteca da assessoria jurídica do Banco do Brasil. **Biblionline**, João Pessoa, v. 8, n. esp., p. 122-133, 2012.

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

ARAÚJO JÚNIOR, R. H.; TARAPANOFF, K. Precisão no processo de busca e recuperação da informação: uso da mineração de textos. **Ciência da Informação**, v. 35, n. 3, 2006.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **NBR 12676**: métodos para análise de documentos – determinação de seus assuntos e seleção de termos de indexação. Rio de Janeiro, 1992. 4 p.

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

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.

BELLUZZO, R. C. B. et al. Indexação automática de documentos com "PRECIS Software" no serviço de biblioteca e documentação da Faculdade de Odontologia de Bauru (USP). In: SIMPÓSIO LATINO-AMERICANO DE TERMINOLOGIA, 2., 1990, Brasília. **Anais...** Brasília: IBICT, 1990. p. 159-164.

BOCCATO, V. R. C. A linguagem documentária em catálogos on-line para política de indexação. In: GIL LEIVA, I.; FUJITA, M. S. L. (Ed.). **Política de indexação**. São Paulo: Cultura Acadêmica; Marília: Oficina Universitária, 2012. p. 139-151.

BOCCATO, V. R. C. **Avaliação do uso de linguagem documentária em catálogos coletivos de bibliotecas universitárias**: um estudo sociocognitivo com protocolo verbal. 2009. 301f. Tese (Doutorado) - Universidade Estadual Paulista, campus de Marília. São Paulo, 2009.

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. Indexação automática de documentos textuais: proposta de critérios essenciais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 11., 2010, Rio de Janeiro, RJ. **Anais...**, Rio de Janeiro, RJ: ENANCIB, 2010.

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.

BRÄSCHER, M. A Ambiguidade na Recuperação da Informação, **DataGramZero** - Revista de Ciência da Informação, v. 3, n. 1, fev. 2002.

BRASIL, M. I. et al. Vocabulário sistematizado: a experiência da Fundação Casa de Rui Barbosa. In: FEBAB. **Integrar**: 1º Congresso Internacional de Arquivos, Bibliotecas, Centros de Documentação e Museus, textos. 2002. 12 p.

BRITO, M. Sistemas de informação em linguagem natural: em busca de uma indexação automática, **Ciência da Informação**, Brasília, v. 21, n. 3, p. 223-232, set./dez. 1992.

BRUZINGA, G. S.; MACULAN, B. C. M. S.; LIMA, G. A. B. O. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 8., 2007, Bahia. Anais..., Bahia: ENANCIB, 2007.

CAFÉ, L. Contribuições da Gramática Funcional da delimitação de segmentos descritores de informação. In: RODRIGUES, G. M.; LOPES, I. L. (Org.). **Organização e representação do conhecimento na perspectiva da ciência da informação**. Brasília: Thesaurus, 2003, p. 118-140. (Estudos Avançados em Ciência da Informação, v.2).

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.

CÂMARA JÚNIOR, A. T. Indexação automática de acórdãos por meio de processamento de linguagem natural. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 8., 2007, Salvador. **Anais...**, Salvador: ENANCIB, 2007.

CÂMARA JÚNIOR, A. T. **Processamento de linguagem natural para indexação automática semântico-ontológica**. Brasília, DF, 2013. 180f. Tese (Doutorado) – Departamento de Ciência da Informação e Documentação da Universidade de Brasília. Brasília, DF, 2013.

CANFORA, G.; CERULO, L. Taxonomy of information retrieval models and tools. **Journal of Computing and Information Technology**, v. 12, n. 3, p. 175-194, 2004.

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

CARDOSO, O. N. P. Recuperação de informação. **INFOCOMP Journal of Computer Science**, vol. 2, n. 1, p. 27-32, 2000.

CARNEIRO, M. V. Diretrizes para uma política de indexação. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 14, n. 2, p. 221-241, set. 1985.

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

CAVALCANTI, C. R. Indexação, **Estudos Avançados em Biblioteconomia e Ciência da Informação**, Brasília, v. 1, n. 1, p. 211-233, 1982.

CENDÓN, B. V. Ferramentas de busca na WEB, **Ciência da Informação**, Brasília, v. 30, n. 1, p. 39-49, jan./abr. 2001.

CESARINO, M. A. N. Sistemas de recuperação da informação. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 14, n. 2, p. 157-168, set. 1985.

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.

CHAUMIER, J. Indexação: conceito, etapas e instrumentos. Trad. José Augusto Chaves Guimarães. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v.21, n.1/2, p.63-79, jan./jun. 1988.

CORRÊA, R. F. et al. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais, **AtoZ**, Curitiba, v. 1, n. 1, p. 11-22, jan./jun. 2011.

CUNHA, M. B. **Para saber mais**: fontes de informação em ciência e tecnologia. Brasília: Briquet de Lemos / Livros, 2001.

CUNHA, M. B.; CAVALCANTI, R. O. **Dicionário de biblioteconomia e arquivologia**. Brasília: Briquet de Lemos/Livros, 2008.

DIAS, E. W.; NAVES, M. L. **Análise de assunto:** teoria e prática. 2. ed. rev. Brasília, DF: Briquet de Lemos/Livros, 2013.

ESTUDILLO GARCÍA, Joel. Elementos que conforman la sociedad de la información. **Investigación Bibliotecológica**, v. 15, n. 31, p. 161-194, jul./dic. 2001.

FERNEDA, E. **Recuperação de informação:** análise da contribuição da ciência da computação para a ciência da informação. São Paulo, 2003. 147 p. Tese (Doutorado em Ciência da Informação e Documentação) – Universidade de São Paulo. São Paulo, 2003.

FERNEDA, E. **Introdução aos modelos computacionais de recuperação de informação.** Rio de Janeiro: Editora Ciência Moderna, 2012.

FERNEDA, E.; GALVÃO, M. C. B.; ROCHA, J. E. S. Um Método de Indexação Automática de Documentos: aplicação em laudos de exames radiológicos. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 11., 2010, Rio de Janeiro. **Anais...** Rio de Janeiro: ANCIB, 2010.

FRANCELIN, M. M. Configuração epistemológica da ciência da informação no Brasil em uma perspectiva

pós-moderna: análise de periódicos da área. **Ci. Inf.**, Brasília, v. 33, n. 2, p. 49-66, maio/ago. 2004

FUJITA, M. S. L. Avaliação da eficácia de recuperação do sistema de indexação Precis. **Ci. Inf.**, Brasília, v. 18, n. 2, p. 120-134, jul./dez. 1989.

FUJITA, M. S. L. A identificação de conceitos no processo de análise de assunto para indexação. **Revista Digital de Biblioteconomia e Ciência da Informação**, Campinas, v. 1, n. 1, p. 60-90, jul./dez. 2003.

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.

FUJITA, M. S. L. A leitura documentária na perspectiva de suas variáveis: leitor-contexto. **DataGramaZero**, v. 5, n. 4, ago. 2004.

FUJITA, M. S. L. Linguagem documentária em odontologia: uma aplicação do sistema de indexação "PRECIS". In: SIMPÓSIO LATINO-AMERICANO DE TERMINOLOGIA, 2., 1990, Brasília. **Anais...** Brasília: IBICT, 1990. p. 159-164.

FUJITA, M. S. L. A política de indexação para representação e recuperação da informação. In: GIL LEIVA, I.; FUJITA, M. S. L. (Ed.). **Política de indexação**. São Paulo: Cultura Acadêmica; Marília: Oficina Universitária, 2012. p. 17-28.

FUJITA, M. S. L. **Precis na língua portuguesa**: teoria e prática de indexação. Brasília: Editora Universidade de Brasília, 1989.

FUJITA, M. S. L. A representação documentária no processo de indexação com o modelo de leitura documentária para textos científicos e livros: uma abordagem cognitiva com protocolo verbal. **Ponto de Acesso**, Salvador, v. 7, n. 1, p. 42-66, abr. 2013.

FUJITA, M. S. L.; BOCCATO, V. R. C.; RUBI, M. P.; O contexto da indexação para a catalogação de livros em abordagem sociocognitiva. **Brazilian Journal of Information Science**, Marília (SP), v. 4, n. 2, p. 22-40, jul./dez. 2010.

FUJITA, M. S. L.; RUBI, M. P.; BOCCATO, V. R. C. As diferentes perspectivas teóricas e metodológicas sobre indexação e catalogação de assuntos. In: 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. p. 19-42.

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

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

GIL LEIVA, I. Aspectos conceituais da indexação. In: GIL LEIVA, I.; FUJITA, M. S. L. (Ed.). **Política de indexação**. São Paulo: Cultura Acadêmica; Marília: Oficina Universitária, 2012. p. 31-105.

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.

GIL LEIVA, I.; RODRÍGUEZ MUÑOZ, J. V. Los Orígenes del almacenamiento y recuperación de información. **Boletín de la Asociación Andaluza de Bibliotecarios**, Málaga, n. 42, p. 51-66, 1996.

GIL LEIVA, I.; RODRÍGUEZ MUÑOZ, J. V. Tendencias em los sistemas de indización automática. Estudio evolutivo. **Revista Española de Documentación Científica**, v. 19, n. 3, p. 273-291, 1996.

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

GONZALEZ, M.; LIMA, V. L. S. Recuperação de informação e processamento da linguagem natural. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 23., 2003, Campinas. **Anais da Jornada de Mini-Cursos de Inteligência Artificial**, 3. Campinas: [s.n.], v. 3, p.347-395.

GOTTSCHALG-DUQUE, C. **SiRILiCO**: Uma Proposta para um sistema de recuperação de informação baseado em teorias da lingüística computacional e ontologia. Belo Horizonte, 2005. 120f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação da Universidade Federal de Minas Gerais. Belo Horizonte, 2005.

GUEDES, V. L. S. Estudos de um critério para indexação automática derivativa de textos científicos e tecnológicos, **Ciência da Informação**, 23, n. 3, p. 318-326, set./dez. 1994.

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.

GUINCHAT, C.; MENOU, M. **Introdução geral às ciências e técnicas da informação e documentação**. 2. ed. corr. aum. Brasília: IBICT, 1994.

HABRAN, J. MISTRAL: sistema de recuperação documentária, **Revista de Biblioteconomia de Brasília**, Brasília, v. 13, n. 1, p. 67-71, jan./jun. 1985.

HALLER, J. Análise automática de textos em sistemas de informação, **Revista de Biblioteconomia de Brasília**, Brasília, v. 11, n. 1, p. 105-113, jan./jun. 1983.

HALLER, J. Indexação automática de textos, **Revista de Biblioteconomia de Brasília**, Brasília, v. 13, n. 1, p. 27-32, jan./jun. 1985.

HJØRLAND, B. Automatic Indexing. In: _____. **Lifeboat for Knowledge Organization**. [s.l.]:[s.n.], 2008.

HLAVA, M. M. Automatic indexing: a matter of degree. **Bulletin of the American Society for Information**

Science and Technology, v. 29, n. 1, p. 12-15, out./nov. 2002.

HOLANDA, C.; BRAZ, M. I. Indexação automática de conteúdos na web: análise de sites de museus, **Biblionline**, João Pessoa, v. 8, n. 1, p. 42-59, 2012.

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.

KURAMOTO, H. Proposta de um Sistema de Recuperação de Informação Assistido por Computador – SRIAC, **Revista de Biblioteconomia de Brasília**, Brasília, v. 21, n.2, p. 211-228, jul./dez. 1997.

KURAMOTO, H. Sintagmas Nominais: uma Nova Proposta para a Recuperação de Informação, **DataGramaZero** - Revista de Ciência da Informação, v. 3, n. 1, fev. 2002.

LADEIRA, A. P. **Processamento de Linguagem Natural**: caracterização da produção científica dos pesquisadores brasileiros. 2010. 262f. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais. Belo Horizonte, 2010.

LADEIRA, A. P.; ALVARENGA, L. Processamento de linguagem natural: em busca de evidências temáticas nas publicações nacionais e contemporâneas. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 10., 2009, João Pessoa, PB. Anais..., João Pessoa, PB: ENANCIB, 2009.

LANCASTER, F. W. **Avaliação de serviços de bibliotecas**. Brasília, DF: Briquet de Lemos / Livros, 1996.

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

LARA, Marilda Lopez Ginez de. Diferenças conceituais sobre termos e definições e implicações na organização da linguagem documentária. **Ci. Inf.**, Brasília, v. 33, n. 2, p. 92, maio/ago. 2004.

LIMA, V. M. A.; BOCCATO, V. R. C. O desempenho terminológico dos descritores em ciência da

informação do vocabulário controlado do SIBI/USP nos processos de indexação manual, automática, e semi-automática, *Perspectivas em Ciência da Informação*, v. 14, n. 1, p. 131-151, jan./abr. 2009.

LOPES, E. F. Avaliação de serviços de indexação e resumo: critérios, medidas e metodologia. **R. Esc. Bibliotecon.**, Belo Horizonte, v. 14, n. 2, p. 242-256, set. 1985

LOPES, I. L. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. **Ci. Inf.**, Brasília, v. 31, n. 1, p. 41-52, jan./abr. 2002.

LUHN, H. P. A statistical approach to mechanized encoding and searching of literary information. **IBM Journal of Research and Development**, v. 1, n. 4, p. 309-317, 1957.

MACIAS-CHAPULA, C. A. O papel da informetria e da cienciometria e sua perspectiva nacional e internacional. **Ciência da Informação**, Brasília, v. 27, n. 2, p. 134-140, maio/ago. 1998.

MACIEL, A. C.; MENDONÇA, M. A. R. **Bibliotecas como organizações**. 1. ed. rev. Rio de Janeiro: Interciência; Niterói: Intertexto, 2006.

MAI, J-E. Analysis in indexing: document and domain centered approaches. **Information Processing and Management**: an international journal, Nova York, v. 41, p. 599-611, feb. 2005.

MAI, J-E. Deconstructing the indexing process, **Advances in Librarianship**, v. 23, p. 269-298, 2000.

MAI, J-E. Semiotics and indexing: an analysis of the subject indexing process. **Journal of Documentation**, v. 57, n. 5, p. 591-622, 2001.

MAIA, E. L. S. Comportamento bibliométrico da língua portuguesa, como veículo de representação da informação. **Ciência da Informação**, Rio de Janeiro, v.2, n.2, p.99-138, 1973.

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 - UFMG. Belo Horizonte, 2008.

MAIA, L. C. G.; SOUZA, R. R. Uso de sintagmas nominais na classificação automática de documentos eletrônicos, **Perspectivas em Ciência da Informação**, v. 15, n. 1, p. 154-172, jan./abr. 2010.

MAMFRIM, F. P. B. Representação de conteúdo via indexação automática em textos integrais em língua portuguesa, **Ciência da Informação**, Brasília, v. 20, n. 2, p. 191-203, jul./dez. 1991.

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.

MARQUES JÚNIOR, A. M. Indexação automática de atas: a experiência no Curso de Pós-Graduação em Biblioteconomia da UFMG, **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 20, n. 1, p. 100-129, jan./jun. 1991.

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.

MENDES, F. G. **Do pergaminho ao texto eletrônico**: evolução das tecnologias de leitura e escrita. Projeto A leitura no aperfeiçoamento do ensino: uma proposta de pesquisa participante, Rio de Janeiro, dez. 2008.

MENDES, M. T. P.; SIMÕES, M. da G. **Indexação por assuntos**: princípios gerais e normas. Lisboa: Gabinete de Estudos a&b, 2002. (Estudos a&b; 1).

MENDONÇA, E. S. A linguística e a ciência da informação: estudos de uma interseção, **Ciência da Informação**, Brasília, v. 29, n. 3, p. 50-70, set./dez. 2000.

MIGUÉIS, A.; NEVES, B. Uma abordagem à linguagem de indexação dos artigos científicos depositados no repositório científico da Universidade de Coimbra. **Ponto de Acesso**, Salvador, v.7, n.1, p. 116-131, 2013.

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.

MORIN, E. **Ciência com consciência**. 15.ed. rev. e modificada. Rio de Janeiro: Bertrand Brasil, 2013. 344 p.

MOURA, L. T. T. C. Indexação automática derivativa: estudo exploratório em citações comparativamente a

resumos, Rio de Janeiro, 1995. 406f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro. Rio de Janeiro, 1995.

MOURA, M. A. et al. Linguagens de indexação em contextos cinematográficos: a experiência de elaboração do tesauro eletrônico do cinema brasileiro. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 10, n. 1, p. 54-69, jan./jun. 2005.

NARUKAWA, C. M. **Estudo de Vocabulário Controlado na Indexação Automática: Aplicação no Processo de Indexação do Sistema de Indización Semiautomática (SISA)**. 2011. 222 f. Dissertação (Mestrado) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2011.

NARUKAWA, C. M.; FUJITA, M. S. L. O uso de linguagens documentárias alfabéticas na indexação automatizada. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 16., 2010, Rio de Janeiro. Anais..., Rio de Janeiro: ENANCIB, 2010.

NARUKAWA, C. M.; GIL LEIVA, I.; FUJITA, M. S. L. Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software SISA com uso da terminologia DeCS na área de odontologia. **Informação & Sociedade: Estudos**, João Pessoa, v.19, n.2, p. 99-118, maio/ago. 2009.

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.

NAVES, M. M. L. **Curso de indexação**: princípios e técnicas de indexação, com vistas à recuperação da informação. Belo Horizonte: UFMG, Biblioteca Universitária, 2004. Material didático.

OLIVEIRA, E. et al. Um modelo algébrico para representação, indexação e classificação automática de documentos digitais, *Revista Brasileira de Biblioteconomia e Documentação*, Nova Série, São Paulo, v.3, n.1, p.73-98, jan./jun. 2007.

PINHEIRO, L. V. R. Medidas de consistência da indexação: interconsistência. **Ciência da Informação**, Rio de Janeiro, v.7, n.2, p.109-114, 1978.

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

PINTO, M. C. M. F. Análise e representação de assuntos em sistemas de recuperação de informação; linguagens de indexação. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 14, n. 2, p. 169-186, set. 1985.

PINTO, V. B. Indexação documentária: uma forma de representação do conhecimento registrado, **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 6, n. 2, p. 223-234, jul./dez. 2001.

QUONIAM, L.; BALME, F.; ROSTAMG, H.; GIRAUD, E.; DOU, J. M. Bibliometric law used for information retrieval. **Scientometrics**, 41 (1-2), p. 83-91, 1998.

RIVIER, A. Construção de linguagens de indexação. **Revista da Escola de Biblioteconomia da UFMG**, v. 21, n. 1, p. 56-99, jan. / jun. 1992.

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

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

ROBREDO, J. Otimização dos processos de indexação dos documentos e de recuperação da informação mediante o uso de instrumentos de controle terminológico, **Ciência da Informação**, Brasília, v. 11, n. 1, p. 3-18, 1982.

ROBREDO, J.; CUNHA, M. B. Aplicação de técnicas infométricas para identificar a abrangência do léxico básico que caracteriza os processos de indexação e recuperação da informação, **Ciência da Informação**, Brasília, v. 27, n. 1, p. 11-27, jan./abr. 1998.

ROBREDO, J.; FERREIRA, J. A. d. P. Conceituação de um programa para indexação automática de textos, **Revista de Biblioteconomia de Brasília**, Brasília, v. 8, n. 2, p. 254-263, 1980.

ROCHA, R. P. Metadados, Web Semântica, Categorização Automática: combinando esforços humanos e computacionais para a descoberta e uso dos recursos da web, **Em Questão**, Porto Alegre, v. 10, n. 1, p. 109-121, jan./jun. 2004.

ROWLEY, J. **A biblioteca eletrônica**. 2. ed. Brasília: Brinquet de Lemos, 2002.

RUBI, M. P. Política de indexação. In: GIL LEIVA, I.; FUJITA, M. S. L. (Ed.). **Política de indexação**. São Paulo: Cultura Acadêmica; Marília: Oficina Universitária, 2012. p. 107-120.

RUBI, M. P.; FUJITA, M. S. L. Elementos de política de indexação em manuais de indexação de sistemas de informação especializados. **Perspect. cienc. inf.**, Belo Horizonte, v. 8, n. 1, p.66-77, jan./jun. 2003.

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, SP: Editora Átomo, 2003.

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.

SAYÃO, F. Bases de dados: a metáfora da memória científica. **Ci. Inf.**, Brasília, v. 25, n. 3, p. 314-318, set./dez. 1996.

SILVA, A. M.; SILVA, I. R.; ARANTES, L. H. M. Biblioteca digital de peças teatrais, **Ciência da Informação**, Brasília, v. 33, n. 2, p. 187-196, maio/ago. 2004.

SILVA, L. A. G. Políticas e programas de informação e documentação da UNESCO e fontes para seu estudo. **Inf. & Soc.:**Est, João Pessoa, v.4, n.1, p.68-84, jan./dez. 1994.

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.

SILVA, W. T.; MILIDIÚ, R. L. Indexação e recuperação da informação com função de crença, **Ciência da Informação**, Brasília, v. 20, n. 2, p. 155-164, jul./dez. 1991.

SOERGEL, D. Indexing and retrieval performance: the logical evidence. **Journal of the American Society for Information Science**, Washington, v. 45, n. 8, p. 589-599, 1994.

SOUZA, R. R. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. Belo Horizonte, 2005. 215f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação da Universidade Federal de Minas Gerais. Belo Horizonte, 2005.

SOUZA, R. R. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 6., 2005, Florianópolis, SC. **Anais...**, Florianópolis, SC: ENANCIB, 2005.

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.

SOUZA, R. R.; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação, **Ciência da Informação**, Brasília, v. 33, n. 1, p. 132-141, jan./abril 2004.

STEVENS, M. E. **Automatic indexing**: a state-of-the-art report. 1965. 298f. Monograph 91 – National Bureau of Standards, Washington, D.C., 1970.

STREHL, L. Avaliação da consistência da indexação realizada em uma biblioteca universitária de artes. **Ci. Inf.**, Brasília, v. 27, n. 3, p. 329-335, set./dez. 1998.

UNISIST. Princípios de indexação. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v.10, n.1, p.83-94, mar. 1981.

UNIVERSIDADE FEDERAL DE PERNAMBUCO. Programa de Pós-Graduação em Ciência da Informação. **Linhas de pesquisa**, 2009. Disponível em: <http://www.ufpe.br/ppgci/index.php?option=com_content&view=article&id=302&Itemid=230>. Acesso em: 27 mar 2014.

VIEIRA, S. B. Análise comparativa entre indexação automática e manual da literatura brasileira de ciência da informação. **R. Bibliotecon.**, Brasília, v. 16, n. 1, p. 83-94, jan./jun. 1988.

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

APÊNDICE A – Foco de análise para o panorama ordenada cronologicamente

MAIA, E. L. S. Comportamento bibliométrico da língua portuguesa, como veículo de representação da informação. **Ciência da Informação**, Rio de Janeiro, v.2, n.2, p.99-138, 1973.

ROBREDO, J.; FERREIRA, J. A. d. P. Conceituação de um programa para indexação automática de textos, **Revista de Biblioteconomia de Brasília**, Brasília, v. 8, n. 2, p. 254-263, 1980.

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.

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

ROBREDO, J. Otimização dos processos de indexação dos documentos e de recuperação da informação mediante o uso de instrumentos de controle terminológico, **Ciência da Informação**, Brasília, v. 11, n. 1, p. 3-18, 1982.

CAVALCANTI, C. R. Indexação, **Estudos Avançados em Biblioteconomia e Ciência da Informação**, v. 1, n. 1, p. 211-233, 1982.

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.

HALLER, J. Análise automática de textos em sistemas de informação, **Revista de Biblioteconomia de Brasília**, Brasília, v. 11, n. 1, p. 105-113, jan./jun. 1983.

HABRAN, J. MISTRAL: sistema de recuperação documentária, **Revista de Biblioteconomia de Brasília**, Brasília, v. 13, n. 1, p. 67-71, jan./jun. 1985.

HALLER, J. Indexação automática de textos, **Revista de Biblioteconomia de Brasília**, Brasília, v. 13, n. 1, p. 27-32, jan./jun. 1985.

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

VIEIRA, S. B. Análise comparativa entre indexação automática e manual da literatura brasileira de ciência da informação. **Revista de Biblioteconomia de Brasília**, Brasília, v. 16, n. 1, p. 83-94, jan./jun. 1988.

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

FUJITA, M. S. L. **Precis na língua portuguesa**: teoria e prática de indexação. Brasília: Editora Universidade de Brasília, 1989.

FUJITA, M. S. L. Avaliação da eficácia de recuperação do sistema de indexação PRECIS, **Ciência da Informação**, Brasília, v. 18, n. 2, p. 120-134, jul./dez. 1989.

FUJITA, M. S. L. Linguagem documentária em odontologia: uma aplicação do sistema de indexação "PRECIS". In: SIMPÓSIO LATINO-AMERICANO DE TERMINOLOGIA, 2., 1990, Brasília. **Anais...** Brasília: IBICT, 1990. p. 159-164.

BELLUZZO, R. C. B. et al. Indexação automática de documentos com "PRECIS Software" no serviço de biblioteca e documentação da Faculdade de Odontologia de Bauru (USP). In: SIMPÓSIO LATINO-AMERICANO DE TERMINOLOGIA, 2., 1990, Brasília. **Anais...** Brasília: IBICT, 1990. p. 159-164.

MAMFRIM, F. P. B. Representação de conteúdo via indexação automática em textos integrais em língua portuguesa, **Ciência da Informação**, Brasília, v. 20, n. 2, p. 191-203, jul./dez. 1991.

MARQUES JÚNIOR, A. M. Indexação automática de atas: a experiência no Curso de Pós-Graduação em Biblioteconomia da UFMG, **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 20, n. 1, p. 100-129, jan./jun. 1991.

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.

SILVA, W. T. Indexação e recuperação da informação com função de crença, **Ciência da Informação**, Brasília, v. 20, n. 2, p. 155-164, jul./dez. 1991.

BRITO, M. Sistemas de informação em linguagem natural: em busca de uma indexação automática, **Ciência da Informação**, Brasília, v. 21, n. 3, p. 223-232, set./dez. 1992.

RIVER, Alexis. Construção de linguagens de indexação: aspectos teóricos. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 21, n. 1, p. 56-99, jan./jun. 1992.

GUEDES, V. L. S. Estudos de um critério para indexação automática derivativa de textos científicos e tecnológicos, **Ciência da Informação**, 23, n. 3, p. 318-326, set./dez. 1994.

MOURA, Lilia Teresa Torres Cursino de. Indexação automática derivativa: estudo exploratório em citações comparativamente a resumos, Rio de Janeiro, 1995. 406f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal do Rio de Janeiro. Rio de Janeiro, 1995.

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.

KURAMOTO, H. Proposta de um Sistema de Recuperação de Informação Assistido por Computador – SRIAC, **Revista de Biblioteconomia de Brasília**, Brasília, v. 21, n.2, p. 211-228, jul./dez. 1997.

ROBREDO, J.; CUNHA, M. B. Aplicação de técnicas infométricas para identificar a abrangência do léxico básico que caracteriza os processos de indexação e recuperação da informação, **Ciência da Informação**, Brasília, v. 27, n. 1, p. 11-27, jan./abr. 1998.

MENDONÇA, E. S. A linguística e a ciência da informação: estudos de uma interseção, **Ciência da Informação**, Brasília, v. 29, n. 3, p. 50-70, set./dez. 2000.

ALCAIDE, G. S. et. al. Análise comparativa e de consistência entre representações automática e manual de informações documentárias, **Transinformação**, v. 13, n. 1, p. 23-41, jan./jun. 2001.

CENDÓN, B. V. Ferramentas de busca na WEB, **Ciência da Informação**, Brasília, v. 30, n. 1, p. 39-49, jan./abr. 2001.

PINTO, V. B. Indexação documentária: uma forma de representação do conhecimento registrado, **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 6, n. 2, p. 223-234, jul./dez. 2001.

KURAMOTO, H. Sintagmas Nominais: uma Nova Proposta para a Recuperação de Informação, **DataGramaZero** - Revista de Ciência da Informação, v. 3, n. 1, fev. 2002.

BRÄSCHER, M. A Ambiguidade na Recuperação da Informação, **DataGramaZero** - Revista de Ciência da Informação, v. 3, n. 1, fev. 2002.

CAFÉ, L. Contribuições da Gramática Funcional da delimitação de segmentos descritores de informação. In: RODRIGUES, G. M.; LOPES, I. L. (Org.).

Organização e representação do conhecimento na perspectiva da ciência da informação. Brasília: Thesaurus, 2003, p. 118-140. (Estudos Avançados em Ciência da Informação, v.2).

ROCHA, R. P. Metadados, Web Semântica, Categorização Automática: combinando esforços humanos e computacionais para a descoberta e uso dos recursos da web, **Em Questão**, Porto Alegre, v. 10, n. 1, p. 109-121, jan./jun. 2004.

SILVA, A. M.; SILVA, I. R.; ARANTES, L. H. M. Biblioteca digital de peças teatrais, **Ciência da Informação**, Brasília, v. 33, n. 2, p. 187-196, maio/ago. 2004.

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.; ALVARENGA, L. A Web Semântica e suas contribuições para a ciência da informação, **Ciência da Informação**, Brasília, v. 33, n. 1, p. 132-141, jan./abril 2004.

GOTTSCHALG-DUQUE, C. **SiRILiCO**: Uma Proposta para um sistema de recuperação de informação baseado em teorias da lingüística computacional e ontologia. Belo Horizonte, 2005. 120f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação da Universidade Federal de Minas Gerais. Belo Horizonte, 2005.

SOUZA, R. R. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. Belo Horizonte, 2005. 215f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação da Universidade Federal de Minas Gerais. Belo Horizonte, 2005.

SOUZA, R. R. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 6., 2005, Florianópolis, SC. **Anais...**, Florianópolis, SC: ENANCIB, 2005.

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.

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.

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

BRUZINGA, G. S.; MACULAN, B. C. M. S.; LIMA, G. A. B. O. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 8., 2007, Bahia. **Anais...**, Bahia: ENANCIB, 2007.

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.

CÂMARA JÚNIOR, A. T. Indexação automática de acórdãos por meio de processamento de linguagem natural. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 8., 2007, Bahia. **Anais...**, Bahia: ENANCIB, 2007.

OLIVEIRA, E.; CIARELLI, P. M.; SANTOS, M. H.; COSTA, B. O. Um modelo algébrico para representação, indexação e classificação automática

de documentos digitais, **Revista Brasileira de Biblioteconomia e Documentação**, Nova Série, São Paulo, v.3, n.1, p.73-98, jan./jun. 2007.

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.

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 - UFMG. Belo Horizonte, 2008.

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.

LADEIRA, A. P.; ALVARENGA, L. Processamento de linguagem natural: em busca de evidências temáticas nas publicações nacionais e contemporâneas. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 10., 2009, João Pessoa, PB. **Anais...**, João Pessoa, PB: ENANCIB, 2009.

LIMA, V. M. A.; BOCCATO, V. R. C. O desempenho terminológico dos descritores em ciência da informação do vocabulário controlado do SIBI/USP nos processos de indexação manual, automática, e semi-automática, **Perspectivas em Ciência da Informação**, v. 14, n. 1, p. 131-151, jan./abr. 2009.

NARUKAWA, C. M.; GIL LEIVA, I.; FUJITA, M. S. L. Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software SISA com uso da terminologia DeCS na área de odontologia. **Informação & Sociedade: Estudos**, João Pessoa, v.19, n.2, p. 99-118, maio/ago. 2009.

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. São Paulo, 2009.

BORGES, G. S. B. Indexação automática de documentos textuais: proposta de critérios essenciais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 11., 2010, Rio de Janeiro, RJ. **Anais...**, Rio de Janeiro, RJ: ENANCIB, 2010.

LADEIRA, A. P. **Processamento de Linguagem Natural**: caracterização da produção científica dos pesquisadores brasileiros. 2010. 262f. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais. Belo Horizonte, 2010.

MAIA, L. C. G.; SOUZA, R. R. Uso de sintagmas nominais na classificação automática de documentos eletrônicos, **Perspectivas em Ciência da Informação**, v. 15, n. 1, p. 154-172, jan./abr. 2010.

FERNEDA, E.; GALVÃO, M. C. B.; ROCHA, J. E. S. Um Método de Indexação Automática de Documentos: aplicação em laudos de exames radiológicos. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 11., 2010, Rio de Janeiro. **Anais...**, Rio de Janeiro: ANCIB, 2010.

NARUKAWA, C. M.; FUJITA, M. S. L. O uso de linguagens documentárias alfabéticas na indexação automatizada. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 16., 2010, Rio de Janeiro. **Anais...**, Rio de Janeiro: ENANCIB, 2010.

CORRÊA, R. F.; MIRANDA, D. G.; LIMA, C. O. A.; SILVA, T. J. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais, **AtoZ**, Curitiba, v. 1, n. 1, p. 11-22, jan./jun. 2011.

NARUKAWA, C. M. **Estudo de Vocabulário Controlado na Indexação Automática**: Aplicação no Processo de Indexação do Sistema de Indización Semiautomática (SISA). 2011. 222 f. Dissertação (Mestrado) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2011.

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.

FERNEDA, E. **Introdução aos modelos computacionais de recuperação de informação.** Rio de Janeiro: Editora Ciência Moderna, 2012.

HOLANDA, C.; BRAZ, M. I. Indexação automática de conteúdos na web: análise de sites de museus, **Biblionline**, João Pessoa, v. 8, n. 1, p. 42-59, 2012.

APÊNDICE B – Categorias para análise bibliométrica e conteúdo

Legenda

ID: número de identificação do documento

AUT: nome do(s) responsável(is) intelectual(is) da obra

TIT: título da obra

IPUB: nome da fonte publicadora da obra

ANO: ano em que a obra foi publicada

FONT: tipo de fonte onde a obra foi publicada

INST: instituição(ões) responsável(is) pelo(s) desenvolvimento(s) da obra

CIT: citação dos r responsável(is) intelectual(is) da obra

OOBJ: objetivo da obra

OPRO: quando propõe determinar se trata de um sistema, método ou fórmula

OAPL: quando aplica determinar se trata de um sistema, método ou fórmula

OSMF: nome do sistema, método ou fórmula

MAV: como ocorreu a avaliação do sistema, método ou fórmula

MNC: natureza da obra

MTC: tipologia da obra

MVAL: tipo de indexação utilizada para validar os termos

MIT: se identificação dos termos ocorreu por extração ou atribuição

MED: tratamento do texto na entrada de dados

MNLI: tipologia da natureza da linguagem de indexação

MNTE: tipologia da natureza dos termos

MMAE: método de Identificação/Ponderação/Seleção de termos da análise estatística

MMAL: método de Identificação/Ponderação/Seleção de termos da análise linguística

MMAM: método de Identificação/Ponderação/Seleção de termos do aprendizado de máquina

MMRC: método de Identificação/Ponderação/Seleção de termos da representação do conhecimento

ID: 1

AUT: Elza de Lima e Silva Maia

TIT: Comportamento bibliometrico da lingua portuguesa, como veiculo de representacao da informacao

IPUB: Ciencia da Informacao

ANO: 1973

FONT: Periodico

INST: IBICT

CIT: MAIA, E. L. S.

OOBJ: Aplica

OPRO:

OAPL: Formula

OSMF: leisdezipf ; pontotdegoffman

MAV: Julgamentodoautor

MNC: Textocompleto

MTC: Trabalhocientifico

MVAL: Assistida

MIT: Extracao

MED: Estruturado

MNLI: Natural

MNTE: Palavraisoladas ; termoscompostos

MMAE: Localizacao ; frequencia

MMAL:

MMAM:

MMRC:

ID: 2

AUT: Jaime Robredo ; Jose Adalberto de Paula Ferreira

TIT: Conceituacao de um programa para indexacao Automatica de textos

IPUB: Revista de Biblioteconomia de Brasilia

ANO: 1980

FONT: Periodico

INST: UNB ; BINAGRI

CIT: ROBREDO, J. ; FERREIRA, J. A. P.

OOBJ: Propoe

OPRO: Sistema

OAPL:

OSMF: AUTOMINDEX/II

MAV: Indexacaomanual

MNC: Resumo

MTC: Trabalhocientifico

MVAL: Semiautomatica

MIT: Extracao

MED: Estruturado

MNLI: Natural ; controlada

MNTE: Palavraisoladas ; termoscompostos

MMAE: Radicalizacao ; stopwords ; localizacao ; frequencia

MMAL:

MMAM:

MMRC: vocabulariocontrolado

ID: 3

AUT: Maria Augusta da Nobrega Cesarino ; Maria Cristina Mello Ferreira Pinto

TIT: Analise de assunto

IPUB: Revista de Biblioteconomia de Brasilia

ANO: 1980

FONT: Periodico

INST: UFMG

CIT: CESARINO, M. A. N. ; PINTO, M. C. M. F.

OOBJ: Revisao

OPRO:

OAPL:

OSMF:

MAV:

MNC:

MTC:

MVAL:

MIT:

MED:

MNLI:

MNTE:

MMAE:

MMAL:

MMAM:

MMRC:

ID: 4

AUT: Jaime Robredo

TIT: Otimizacao dos processos de indexacao dos documentos e de recuperacao da informacao mediante o uso de instrumentos de controle terminologico

IPUB: Ciencia da Informacao

ANO: 1982

FONT: Periodico

INST: UNB

CIT: ROBREDO, J.

OOBJ: Aplica

OPRO:

OAPL: Sistema

OSMF: AUTOMINDEX/II

MAV: Indexacaomanual

MNC: Textocompleto
 MTC: Trabalhocientifico
 MVAL: Semiautomatica
 MIT: Atribuicao
 MED: Estruturado
 MNLI: Natural ; controlada
 MNTE: Palavrasisoladas ; termoscompostos
 MMAE: Radicalizacao ; stopwords ; localizacao ; frequencia
 MMAL:
 MMAM:
 MMRC: vocabulariocontrolado

ID: 5
 AUT: Jaime Robredo
 TIT: A indexacao automatica de textos: o presente ja entrou no futuro
 IPUB: Estudos Avancados em Biblioteconomia e Ciencia da Informacao
 ANO: 1982
 FONT: Periodico
 INST: UNB
 CIT: ROBREDO, J.
 OOBJ: Aplica
 OPRO:
 OAPL: Sistema
 OSMF: AUTOMINDEX/II
 MAV: Indexacaomanual
 MNC: Resumo
 MTC: Trabalhocientifico
 MVAL: Semiautomatica
 MIT: Atribuicao
 MED: Estruturado
 MNLI: Natural ; controlada
 MNTE: Palavrasisoladas ; termoscompostos
 MMAE: Radicalizacao ; stopwords ; localizacao ; frequencia
 MMAL:
 MMAM:
 MMRC: vocabulariocontrolado

ID: 6
 AUT: Cordelia R. Cavalcanti
 TIT: Indexacao
 IPUB: Estudos Avancados em Biblioteconomia e Ciencia da Informacao
 ANO: 1982
 FONT: Periodico
 INST: UNB
 CIT: CAVALCANTI, C. R.
 OOBJ: Revisao

OPRO:
 OAPL:
 OSMF:
 MAV:
 MNC:
 MTC:
 MVAL:
 MIT:
 MED:
 MNLI:
 MNTE:
 MMAE:
 MMAL:
 MMAM:
 MMRC:

ID: 7

AUT: Alexandre Andreewski ; Vitoriano Ruas
 TIT: Indexacao automatica baseada em metodos linguisticos e estatisticos
 e sua aplicabilidade a lingua portuguesa
 IPUB: Ciencia da Informacao
 ANO: 1983
 FONT: Periodico
 INST: UNIVERSITE DE PARIS SUD ; PUC-Rio
 CIT: ANDREEWSKI, A. ; RUAS, V.
 OOBJ: Propoe
 OPRO: Sistema
 OAPL:
 OSMF: SPIRIT
 MAV:
 MNC:
 MTC:
 MVAL: Automatica
 MIT: Extracao
 MED: Marcado
 MNLI: Natural ; controlada
 MNTE: Palavrasisoladas ; termoscompostos
 MMAE: Radicalizacao ; stopwords ; pesonumerico ; frequencia
 MMAL: Morfologica ; sintatica ; semantica
 MMAM: classificacao
 MMRC: vocabulariocontrolado

ID: 8

AUT: Johann Haller
 TIT: Analise automatica de textos em sistemas de informacao
 IPUB: Revista de Biblioteconomia de Brasilia

ANO: 1983
 FONT: Periodico
 INST: UNB
 CIT: HALLER, J.
 OOBJ: Propoe
 OPRO: Sistema
 OAPL:
 OSMF: Programaexperimental
 MAV:
 MNC:
 MTC:
 MVAL: semiautomatica
 MIT: Atribuicao
 MED: Estruturado
 MNLI: Natural
 MNTE: Palavraisoladas ; termoscompostos
 MMAE: Radicalizacao ; stopwords ; frequencia ; pesonumerico
 MMAL: Morfologica ; sintatica ; semantica
 MMAM: classificacao
 MMRC: vocabulariocontrolado

ID: 9
 AUT: Jean Habran
 TIT: MISTRAL: sistema de recuperacao documentaria
 IPUB: Revista de Biblioteconomia de Brasilia
 ANO: 1985
 FONT: Periodico
 INST: ABC BULL S.A. Telematic
 CIT: HABRAN, J.
 OOBJ: Propoe
 OPRO: Sistema
 OAPL:
 OSMF: MISTRAL
 MAV:
 MNC:
 MTC:
 MVAL: semiautomatica ; automatica
 MIT: Atribuicao
 MED: Estruturado
 MNLI: Natural ; controlada
 MNTE: Palavraisoladas ; termoscompostos
 MMAE: Radicalizacao ; stopwords; frequencia
 MMAL: Morfologica ; sintatica ; semantica
 MMAM:
 MMRC: vocabulariocontrolado

ID: 10
AUT: Johann Haller
TIT: Indexacao automatica de textos
IPUB: Revista de Biblioteconomia de Brasilia
ANO: 1985
FONT: Periodico
INST: UNB
CIT: HALLER, J.
OOBJ: Propoe
OPRO: Sistema
OAPL:
OSMF: Programaexperimental
MAV:
MNC:
MTC:
MVAL: semiautomatica
MIT: Atribuicao
MED: Estruturado
MNLI: Natural
MNTE: Palavrasisoladas ; termoscompostos
MMAE: Radicalizacao ; stopwords ; frequencia ; pesonumerico
MMAL: Morfologica ; sintatica ; semantica
MMAM: classificacao
MMRC: vocabulariocontrolado

ID: 11
AUT: Simone Bastos Vieira
TIT: Indexacao automatica e manual: revisao de literatura
IPUB: Ciencia da Informacao
ANO: 1988
FONT: Periodico
INST: UNB
CIT: VIEIRA, S. B.
OOBJ: Revisao
OPRO:
OAPL:
OSMF:
MAV:
MNC:
MTC:
MVAL:
MIT:
MED:
MNLI:
MNTE:
MMAE:

MMAL:
MMAM:
MMRC:

ID: 12

AUT: Simone Bastos Vieira

TIT: Analise comparativa entre indexacao automatica e manual da literatura brasileira de ciencia da informacao

IPUB: Revista de Biblioteconomia de Brasilia

ANO: 1988

FONT: Periodico

INST: UNB

CIT: VIEIRA, S. B.

OOBJ: Aplica

OPRO:

OAPL: Sistema

OSMF: BIB/DIALOGO

MAV: Indexacaomanual

MNC: Resumo ; titulo

MTC: Trabalhocientifico

MVAL: Automatica

MIT: Extracao

MED: Naoestruturado

MNLI: Natural

MNTE: Palavrasisoladas

MMAE: Radicalizacao ; stopwords ; localizacao ; frequencia

MMAL:

MMAM:

MMRC:

ID: 13

AUT: Hagar Espanha Gomes

TIT: O indexador face as novas tecnologias de informacao

IPUB: Transinformacao

ANO: 1989

FONT: Periodico

INST: CNPQ/RJ

CIT: GOMES, H. E.

OOBJ: Revisao

OPRO:

OAPL:

OSMF:

MAV:

MNC:

MTC:

MVAL:

MIT:
 MED:
 MNLI:
 MNTE:
 MMAE:
 MMAL:
 MMAM:
 MMRC:

ID: 14
 AUT: Mariangela Spotti Lopes Fujita
 TIT: Avaliacao da eficacia de recuperacao do sistema de indexacao PRECIS
 IPUB: Ciencia da Informacao
 ANO: 1989
 FONT: Periodico
 INST: UNESP
 CIT: FUJITA, M. S. L.
 OOBJ: Aplicaepropoe
 OPRO:
 OAPL: Sistema
 OSMF: PRECIS
 MAV: Indexacaomanual ; participacaodousuario
 MNC: Textocompleto
 MTC: Livro
 MVAL: Semiautomatica
 MIT: Atribuicao
 MED: Naoestruturado
 MNLI: Natural
 MNTE: Palavrasisoladas ; termoscompostos
 MMAE: Frequencia
 MMAL: Sintatica ; semantica
 MMAM:
 MMRC: Vocabulariocontrolado

ID: 15
 AUT: Mariangela Spotti Lopes Fujita
 TIT: Linguagem documentaria em odontologia: uma aplicacao do sistema de indexacao ?PRECIS?
 IPUB: SIMPOSIO LATINO-AMERICANO DE TERMINOLOGIA
 ANO: 1990
 FONT: Anais...
 INST: UNESP
 CIT: FUJITA, M. S. L.
 OOBJ: Aplicaepropoe
 OPRO:
 OAPL: Sistema

OSMF: PRECIS
 MAV: Indexacaomanual ; participacaodousuario
 MNC: Textocompleto
 MTC: Trabalhocientifico
 MVAL: Semiautomatica
 MIT: Atribuicao
 MED: Naoestruturado
 MNLI: Natural
 MNTE: Palavrasisoladas ; termoscompostos
 MMAE: Frequencia
 MMAL: Sintatica ; semantica
 MMAM:
 MMRC: Vocabulariocontrolado

ID: 16

AUT: Regina Celia Baptista Belluzzo ; Mariangela Spotti Lopes Fujita ;
 Eliane Falcao Tuler Xavier ; Maria Helena Souza Ronchesel ; Valeria
 Cristina Spina Trindade ; Jose Roberto Brejao
 TIT: Indexacao automatica de documentos com ?PRECIS Software? no
 servi co de biblioteca e documentacao da Faculdade de Odontologia de
 Bauru (USP)

IPUB: SIMPOSIO LATINO-AMERICANO DE TERMINOLOGIA

ANO: 1990

FONT: Anais...

INST: USP ; UNESP

CIT: BELLUZZO, R. C. B. ; FUJITA, M. S. L.; XAVIER, E. F. T. ;
 RONCHESEL, M. H. S. ; TRINDADE, V. C. S. ; BREJAO, J. R.

OOBJ: Aplicaepropoe

OPRO:

OAPL: Sistema

OSMF: PRECIS

MAV: Indexacaomanual ; participacaodousuario

MNC: Textocompleto

MTC: Trabalhocientifico

MVAL: Semiautomatica

MIT: Atribuicao

MED: Naoestruturado

MNLI: Natural

MNTE: Palavrasisoladas ; termoscompostos

MMAE: Frequencia

MMAL: Sintatica ; semantica

MMAM:

MMRC: Vocabulariocontrolado

ID: 17

AUT: Flavia Pereira Braga Mamfrim

TIT: Representacao de conteudo via indexacao automatica em textos
integrals em lingua portuguesa
IPUB: Ciencia da Informacao
ANO: 1991
FONT: Periodico
INST: UFRJ
CIT: MAMFRIM, F. P. B.
OOBJ: Aplicaepropoe
OPRO:
OAPL: Formula
OSMF: leisdezipf ; pontotdegoffman
MAV: indexacaomanual
MNC: Textocompleto
MTC: Trabalhocientifico
MVAL: Semiautomatica
MIT: Extracao
MED: Estruturado
MNLI: Natural
MNTE: Palavraisoladas ; termoscompostos
MMAE: Localizacao ; frequencia
MMAL:
MMAM:
MMRC:

ID: 18
AUT: Alaor Messias Marques Junior
TIT: Indexacao automatica de atas: a experiencia no Curso de Pos-
Graduacao em Biblioteconomia da UFMG
IPUB: Revista da Escola de Biblioteconomia da UFMG
ANO: 1991
FONT: Periodico
INST: UFMG
CIT: MARQUES JUNIOR, A. M.
OOBJ: Aplica
OPRO:
OAPL: Sistema
OSMF:
MAV: Indexacaomanual
MNC: Textocompleto
MTC: Ata
MVAL: Semiautomatica
MIT: Atribuicao
MED: Estruturado
MNLI: Natural ; controlada
MNTE: Palavraisoladas ; termoscompostos
MMAE:

MMAL:
 MMAM:
 MMRC: Vocabulariocontrolado

ID: 19
 AUT: Jaime Robredo
 TIT: Indexacao automatica de textos: uma abordagem otimizada e simples
 IPUB: Ciencia da Informacao
 ANO: 1991
 FONT: Periodico
 INST: UNB
 CIT: ROBREDO, J.
 OOBJ: Propoe
 OPRO: Sistema
 OAPL:
 OSMF: AUTOMINDEX/II
 MAV:
 MNC: Resumo ; titulo
 MTC: Trabalhocientifico ; carta
 MVAL: Semiautomatica
 MIT: Atribuicao
 MED: Estruturado
 MNLI: Natural ; controlada
 MNTE: Palavrasisoladas ; termoscompostos
 MMAE: Radicalizacao ; stopwords ; localizacao ; frequencia ;
 Dicionarioderaizes
 MMAL: Morfologica ; sintatica ; semantica
 MMAM:
 MMRC:

ID: 20
 AUT: Wagner Teixeira da Silva ; Ruy Luiz Milidui
 TIT: Indexacao e recuperacao da informacao com funcao de crenca,
 IPUB: Ciencia da Informacao
 ANO: 1991
 FONT: Periodico
 INST: UNB ; PUC-Rio
 CIT: SILVA, W. T. ; MILIDIU, R. L.
 OOBJ: Propoe
 OPRO: Metodo
 OAPL:
 OSMF: Funcaodecrenca
 MAV:
 MNC:
 MTC:
 MVAL: Automatico

MIT: Atribuicao
 MED: Naoestruturado
 MNLI: Natural ; controlada
 MNTE: Palavrasisoladas ; termoscompostos
 MMAE: Frequencia ; Pesonumerico
 MMAL: semantica
 MMAM:
 MMRC: vocabulariocontrolado

ID: 21
 AUT: Marcilio de Brito
 TIT: Sistemas de informacao em linguagem natural: em busca de uma indexacao automatica
 IPUB: Ciencia da Informacao
 ANO: 1992
 FONT: Periodico
 INST: Sebrae ; UNB
 CIT: BRITO, M.
 OOBJ: Aplica
 OPRO:
 OAPL: Metodo
 OSMF: Sintagmasnominais
 MAV: indexacaomanual
 MNC:
 MTC:
 MVAL: Assistida
 MIT: Atribuicao
 MED: Estruturado
 MNLI: Natural
 MNTE: Palavrasisoladas
 MMAE:
 MMAL: Morfologica ; sintatica ; semantica ; sintagmasnominais
 MMAM:
 MMRC:

ID: 22
 AUT: Alexis Rivier
 TIT: Construcao de linguagens de Indexacao: aspectos teoricos
 IPUB: Revista da Escola de Biblioteconomia da UFMG
 ANO: 1992
 FONT: Periodico
 INST: Universite de Lyon
 CIT: RIVER, A.
 OOBJ: Revisao
 OPRO:
 OAPL:

OSMF:
 MAV:
 MNC:
 MTC:
 MVAL:
 MIT:
 MED:
 MNLI:
 MNTE:
 MMAE:
 MMAL:
 MMAM:
 MMRC:

ID: 23

AUT: Vania Lisboa da Silveira Guedes

TIT: Estudos de um criterio para indexacao automatica derivativa de textos
 cientificos e tecnologicos

IPUB: Ciencia da Informacao

ANO: 1994

FONT: Periodico

INST: UFRJ

CIT: GUEDES, V. L. S.

OOBJ: Aplica

OPRO:

OAPL: Formula

OSMF: Leideamplitude

MAV: Indexacaomanual

MNC: Textocompleto

MTC: Notatecnica

MVAL: Automatica

MIT: Extracao

MED: Estruturado

MNLI: Natural

MNTE: Termoscompostos ; palavraisoladas

MMAE: Localizacao ; frequencia

MMAL:

MMAM:

MMRC:

ID: 24

AUT: Lilia Teresa Torres Cursino de Moura

TIT: Indexacao automatica derivativa: estudo exploratorio em cita coes
 comparativamente a resumos

IPUB:

ANO: 1995

FONT: Dissertacao
 INST: CNPq ; IBICT ; UFRJ
 CIT: MOURA, L. T. T. C.
 OOBJ: Aplica
 OPRO:
 OAPL: Metodo
 OSMF: Frequencia de ocorrencia
 MAV: Indexacao manual
 MNC: Resumo ; titulo ; citacao
 MTC: Trabalho cientifico
 MVAL: Semiautomatica
 MIT: Extracao
 MED: Estruturado
 MNLI: Natural
 MNTE: Palavras isoladas ; termos compostos
 MMAE: Radicalizacao ; stopwords ; localizacao ; frequencia
 MMAL:
 MMAM:
 MMRC:

ID: 25
 AUT: Helio Kuramoto
 TIT: Uma abordagem alternativa para o tratamento e a recuperacao de informacao textual : os sintagmas nominais
 IPUB: Ciencia da Informacao
 ANO: 1995
 FONT: Periodico
 INST: IBICT
 CIT: KURAMOTO, H.
 OOBJ: Propoe
 OPRO: Metodo
 OAPL:
 OSMF: Sintagmas nominais
 MAV: Participacao do usuario
 MNC: Texto completo
 MTC: Trabalho cientifico
 MVAL: Assistida
 MIT: Extracao
 MED: Estruturado
 MNLI: Natural
 MNTE: Palavras isoladas
 MMAE:
 MMAL: Sintagmas nominais
 MMAM: Classificacao
 MMRC:

ID: 26
AUT: Helio Kuramoto
TIT: Proposta de um Sistema de Recuperacao de Informacao Assistido por Computador ? SRIAC
IPUB: Revista de Biblioteconomia de Brasilia
ANO: 1997
FONT: Periodico
INST: IBICT
CIT: KURAMOTO, H.
OOBJ: Propoe
OPRO: Metodo
OAPL:
OSMF: SRIAC
MAV:
MNC:
MTC:
MVAL: Assistida
MIT: Extracao
MED: Estruturado
MNLI: Natural
MNTE: Palavrasisoladas
MMAE:
MMAL: Sintagmasnominais
MMAM:
MMRC:

ID: 27
AUT: Jaime Robredo ; Murilo Bastos da Cunha
TIT: Aplicacao de tecnicas infometricas para identificar a abrangencia do lexico basico que caracteriza os processos de indexacao e recuperacao da informacao
IPUB: Ciencia da Informacao
ANO: 1998
FONT: Periodico
INST: UNB
CIT: ROBREDO, J. ; CUNHA, M. B.
OOBJ: Aplica
OPRO:
OAPL: Metodo
OSMF: Frequenciadecoocorrencia
MAV: Indexacaomanual
MNC: Verbetes
MTC: Manuscritos
MVAL: Automatica
MIT: Extracao
MED: Naoestruturado

MNLI: Natural
 MNTE: Palavras isoladas ; termos compostos
 MMAE: Stopwords ; frequencia ; coocorrencia
 MMAL:
 MMAM: Cluster
 MMRC:

ID: 28
 AUT: Ercilia Severina Mendonca
 TIT: A linguística e a ciência da informação: estudos de uma interseção
 IPUB: Ciência da Informação
 ANO: 2000
 FONT: Periódico
 INST: UNIRIO
 CIT: MENDONCA, E. S.
 OOBJ: Revisão
 OPRO:
 OAPL:
 OSMF:
 MAV:
 MNC:
 MTC:
 MVAL:
 MIT:
 MED:
 MNLI:
 MNTE:
 MMAE:
 MMAL:
 MMAM:
 MMRC:

ID: 29
 AUT: Gabriel Santos Alcaide ; Roberto Julio Gava ; Willame Santos Rodrigues ; Debora Ferreira Santos
 TIT: Análise comparativa e de consistência entre representações automáticas e manuais de informações documentárias
 IPUB: Transinformação
 ANO: 2001
 FONT: Periódico
 INST: FESP
 CIT: ALCAIDE, G. S. ; GAVA, R. J. ; RODRIGUES, W. S. ; SANTOS, D. F.
 OOBJ: Aplicação
 OPRO:
 OAPL: Método
 OSMF: Estatística morfológica

MAV: Indexacaomanual
 MNC: Textocompleto ; titulo
 MTC: Trabalhocientifico
 MVAL: Semiautomatica
 MIT: Extracao
 MED: Estruturado
 MNLI: Controlada
 MNTE: Palavrasisoladas
 MMAE: Stopwords ; frequencia
 MMAL: Morfologica
 MMAM:
 MMRC: Vocabulariocontrolado

ID: 30
 AUT: Beatriz Valadares Cendon
 TIT: Ferramentas de busca na WEB
 IPUB: Ciencia da Informacao
 ANO: 2001
 FONT: Periodico
 INST: UFMG
 CIT: CENDON, B. V.
 OOBJ: Revisao
 OPRO:
 OAPL:
 OSMF:
 MAV:
 MNC:
 MTC:
 MVAL:
 MIT:
 MED:
 MNLI:
 MNTE:
 MMAE:
 MMAL:
 MMAM:
 MMRC:

ID: 31
 AUT: Virginia Bentes Pinto
 TIT: Indexacao documentaria: uma forma de Representacao do conhecimento registrado
 IPUB: Perspectivas em Ciencia da Informacao
 ANO: 2001
 FONT: Periodico
 INST: UFC

CIT: PINTO, V. B.

OOBJ: Revisao

OPRO:

OAPL:

OSMF:

MAV:

MNC:

MTC:

MVAL:

MIT:

MED:

MNLI:

MNTE:

MMAE:

MMAL:

MMAM:

MMRC:

ID: 32

AUT: Helio Kuramoto

TIT: Sintagmas Nominais: uma Nova Proposta para a Recuperacao de Informacao

IPUB: DataGramaZero - Revista de Ciencia da Informacao

ANO: 2002

FONT: Periodico

INST: UFMG

CIT: KURAMOTO, H.

OOBJ: Propoe

OPRO: Metodo

OAPL:

OSMF: Sintagmasnominais

MAV:

MNC:

MTC:

MVAL: Assistida

MIT: Extracao

MED: Estruturado

MNLI: Natural

MNTE: Palavrasisoladas

MMAE:

MMAL: Sintagmasnominais

MMAM:

MMRC:

ID: 33

AUT: Marisa Brascher

TIT: Ambiguidade na Recuperacao da Informacao
 IPUB: DataGramaZero - Revista de Ciencia da Informacao
 ANO: 2002
 FONT: Periodico
 INST: IBICT
 CIT: BRASCHER, M.
 OOBJ: Aplica
 OPRO:
 OAPL: Sistema
 OSMF: Zstation
 MAV:
 MNC: Textocompleto
 MTC: Documentosoficiais
 MVAL: Automatica
 MIT: Atribuicao
 MED: Naoestruturado
 MNLI: Natural ; controlada
 MNTE: Palavrasisoladas ; termoscompostos
 MMAE:
 MMAL: Morfologica ; sintatica ; sintagmasnominais ; semantica
 MMAM:
 MMRC: Graficosconceituais ; vocabulariocontrolado

ID: 34
 AUT: Rafael Port da Rocha
 TIT: Metadados, Web Semantica, Categorizacao automatica: combinando
 esforcos humanos e computacionais para a descoberta e uso dos recursos
 da web
 IPUB: Em Questao
 ANO: 2004
 FONT: Periodico
 INST: UFRGS
 CIT: ROCHA, R. P.
 OOBJ: Revisao
 OPRO:
 OAPL:
 OSMF:
 MAV:
 MNC:
 MTC:
 MVAL:
 MIT:
 MED:
 MNLI:
 MNTE:
 MMAE:

MMAL:
MMAM:
MMRC:

ID: 35
AUT: Angela Maria Silva ; Ilmerio Reis Silva ; Luiz Humberto Martins Arantes
TIT: Biblioteca digital de pecas teatrais
IPUB: Ciencia da Informacao
ANO: 2004
FONT: Periodico
INST: UFU ; FAPEMIG
CIT: SILVA, A. M. ; SILVA, I. R. ; ARANTES, L. H. M.
OOBJ: Aplica
OPRO:
OAPL: Metodo
OSMF: Frequenciadeocorrencia
MAV:
MNC: Textocompleto
MTC: Pecasteatrais
MVAL: Assistida
MIT: Atribuicao
MED: Naoestruturado
MNLI: Natural
MNTE: Palavraisoladas ; termoscompostos
MMAE: Frequencia
MMAL:
MMAM:
MMRC:

ID: 36
AUT: Maria dos Remedios da Silva ; Mariangela Spotti Lopes Fujita
TIT: A pratica de Indexacao: analise da evolucao de tendencias teoricas e metodologicas
IPUB: Transinformacao
ANO: 2004
FONT: Periodico
INST: UNESP
CIT: SILVA, M. R. ; FUJITA, M. S. L.
OOBJ: Revisao
OPRO:
OAPL:
OSMF:
MAV:
MNC:
MTC:

MVAL:
MIT:
MED:
MNLI:
MNTE:
MMAE:
MMAL:
MMAM:
MMRC:

ID: 37
AUT: Renato Rocha Souza ; Lidia Alvarenga
TIT: A Web Semantica e suas Contribuicoes para a ciencia da informacao
IPUB: Ciencia da Informacao
ANO: 2004
FONT: Periodico
INST: UFMG
CIT: SOUZA, R. R. ; ALVARENGA, L.
OOBJ: Revisao
OPRO:
OAPL:
OSMF:
MAV:
MNC:
MTC:
MVAL:
MIT:
MED:
MNLI:
MNTE:
MMAE:
MMAL:
MMAM:
MMRC:

ID: 38
AUT: Renato Rocha Souza
TIT: Uma proposta de metodologia para indexacao automatica utilizando
sintagmas nominais
IPUB: ENCONTRO NACIONAL DE PESQUISA EM CIENCIA DA INFORMACAO
ANO: 2005
FONT: Anais...
INST: UFMG
CIT: SOUZA, R. R.
OOBJ: Aplicaepropoe
OPRO:

OAPL: Metodo
 OSMF: Sintagmasnominais
 MAV: Indexacaomanual
 MNC: Textocompleto
 MTC: Trabalhocientifico
 MVAL: Semiautomatica
 MIT: Extracao
 MED: Estruturado
 MNLI: Natural
 MNTE: Palavrasisoladas
 MMAE: Stopwords ; frequencia
 MMAL: Morfologica ; sintatica ; semantica ; sintagmasnominais
 MMAM:
 MMRC:

ID: 39
 AUT: Vania L. S. Guedes ; Suzana Borschiver
 TIT: Bibliometria: uma ferramenta estatistica para a gestao da informacao e do conhecimento, em sistemas de informacao, de comunicacao e de Avaliacao cientifica e tecnologica
 IPUB: ENCONTRO NACIONAL DE CIENCIA DA INFORMACAO
 ANO: 2005
 FONT: Anais...
 INST: UFRJ
 CIT: GUEDES, V. L. S. ; BORSCHIVER, S.
 OOBJ: Revisao
 OPRO:
 OAPL:
 OSMF:
 MAV:
 MNC:
 MTC:
 MVAL:
 MIT:
 MED:
 MNLI:
 MNTE:
 MMAE:
 MMAL:
 MMAM:
 MMRC:

ID: 40
 AUT: Renato Rocha Souza
 TIT: Uma proposta de metodologia para escolha automatica de descritores utilizando sintagmas nominais

IPUB:
 ANO: 2005
 FONT: Tese
 INST: UFMG
 CIT: SOUZA, R. R.
 OOBJ: Aplicaepropoe
 OPRO:
 OAPL: Metodo
 OSMF: Sintagmasnominais
 MAV: Indexacaomanual
 MNC: Textocompleto
 MTC: Trabalhocientifico
 MVAL: Semiautomatica
 MIT: Extracao
 MED: Estruturado
 MNLI: Natural
 MNTE: Palavrasisoladas
 MMAE: Stopwords ; frequencia
 MMAL: Morfológica ; sintática ; semântica ; sintagmasnominais
 MMAM:
 MMRC:

ID: 41
 AUT: Renato Rocha Souza
 TIT: Uma proposta de metodologia para indexacao automatica utilizando
 sintagmas nominais
 IPUB: Encontros Bibli: Revista Eletronica de Biblioteconomia e Ciencia da
 Informacao
 ANO: 2006
 FONT: Periodico
 INST: UFRJ
 CIT: SOUZA, R. R.
 OOBJ: Aplicaepropoe
 OPRO:
 OAPL: Metodo
 OSMF: Sintagmasnominais
 MAV: Indexacaomanual
 MNC: Textocompleto
 MTC: Trabalhocientifico
 MVAL: Semiautomatica
 MIT: Extracao
 MED: Estruturado
 MNLI: Natural
 MNTE: Palavrasisoladas
 MMAE: Stopwords ; frequencia
 MMAL: Morfológica ; sintática ; semântica ; sintagmasnominais

MMAM:

MMRC:

ID: 42

AUT: Graciane Silva Bruzinga ; Benildes Coura Moreira dos Santos Maculan ; Gercina Angela Borem de Oliveira Lima

TIT: Indexacao Automatica e semantica: estudo da analise do conteudo de teses e dissertacoes

IPUB: ENCONTRO NACIONAL DE PESQUISA EM CIENCIA DA INFORMACAO,

ANO: 2007

FONT: Anais...

INST: UFMG

CIT: BRUZINGA, G. S.; MACULAN, B. C. M. S.; LIMA, G. A. B. O.

OOBJ: Propoe

OPRO: Metodo

OAPL:

OSMF: Sintagmasnominais

MAV:

MNC: Textocompleto

MTC: Trabalhocientifico

MVAL: Automatica

MIT: Atribuicao

MED: Naoestruturado

MNLI: Natural ; controlada

MNTE: Palavrasisoladas

MMAE:

MMAL: Morfologica ; sintatica ; semantica ; sintagmasnominais

MMAM:

MMRC: Vocabulariocontrolado

ID: 43

AUT: Auto Tavares da Camara Junior

TIT: Indexacao automatica de acordaos por meio de processamento de linguagem natural

IPUB: ENCONTRO NACIONAL DE PESQUISA EM CIENCIA DA INFORMACAO

ANO: 2007

FONT: Anais...

INST: UNB

CIT: CAMARA JUNIOR, A. T.

OOBJ: Aplicaepropoe

OPRO:

OAPL: Metodo

OSMF: Sintagmasnominais

MAV: Indexacaomanual

MNC: Textocompleto

MTC: Acordao

MVAL: Automatica
 MIT: Atribuicao
 MED: Naoestruturado
 MNLI: Natural ; controlada
 MNTE: Palavraisoladas ; termoscompostos
 MMAE: pesonumerico
 MMAL: Morfologica ; sintatica ; semantica ; sintagmasnominais
 MMAM: Cluster
 MMRC: vocabulariocontrolado

ID: 44

AUT: Elias Oliveira ; Patrick Marques Ciarelli ; Marcos Hercules Santos ; Bruno Oliveira da Costa
 TIT: Um modelo algebrico para Representacao, indexacao e classificacao automatica de documentos digitais
 IPUB: Revista Brasileira de Biblioteconomia e Documentacao
 ANO: 2007
 FONT: Periodico
 INST: UFES
 CIT: OLIVEIRA, E. ; CIARELLI, P. M. ; SANTOS, M. H. ; COSTA, B. O.
 OOBJ: Aplica
 OPRO:
 OAPL: Metodo
 OSMF: Frequenciadeocorrencia
 MAV: Indexacaomanual
 MNC: Textocompleto
 MTC: noticiaRSS
 MVAL: Semiautomatica ; automatica
 MIT: Extracao
 MED: Naoestruturado
 MNLI: Natural
 MNTE: Palavraisoladas
 MMAE: Stopwords ; frequencia ; pesonumerico
 MMAL:
 MMAM:
 MMRC:

ID: 45

AUT: Juliano Serra Barreto
 TIT: Desafios e avancos na recuperacao automatica da informacao audiovisual
 IPUB: Ciencia da Informacao
 ANO: 2007
 FONT: Periodico
 INST: UNB
 CIT: BARRETO, J. S.

OOBJ: Revisao

OPRO:

OAPL:

OSMF:

MAV:

MNC:

MTC:

MVAL:

MIT:

MED:

MNLI:

MNTE:

MMAE:

MMAL:

MMAM:

MMRC:

ID: 46

AUT: Auto Tavares da Camara Junior

TIT: Indexacao automatica de acordaos por meio de processamento de linguagem natural

IPUB:

ANO: 2007

FONT: Dissertacao

INST: UNB

CIT: CAMARA JUNIOR, A. T.

OOBJ: Aplicaepropoe

OPRO:

OAPL: Metodo

OSMF: Sintagmasnominais

MAV: Indexacaomanual

MNC: Textocompleto

MTC: Acordao

MVAL: Automatica

MIT: Atribuicao

MED: Naoestruturado

MNLI: Natural ; controlada

MNTE: Palavraisoladas ; termoscompostos

MMAE: pesonumerico

MMAL: Morfologica ; sintatica ; semantica ; sintagmasnominais

MMAM:

MMRC: vocabulariocontrolado

ID: 47

AUT: Graciane Silva Bruzinga Borges ; Benildes Coura Moreira dos Santos Maculan ; Gercina Angela Borem de Oliveira Lima

TIT: Indexacao Automatica e semantica: estudo de analise do conteudo de teses e dissertacoes
 IPUB: Informacao & Sociedade: Estudos
 ANO: 2008
 FONT: Periodico
 INST: UFMG
 CIT: BORGES, G. S. B. ; MACULAN, B. C. M. S. ; LIMA, G. A. B. O.
 OOBJ: Propoe
 OPRO: Metodo
 OAPL:
 OSMF: Sintagmasnominais
 MAV:
 MNC: Textocompleto
 MTC: Trabalhocientifico
 MVAL: Automatica
 MIT: Atribuicao
 MED: Naoestruturado
 MNLI: Natural ; controlada
 MNTE: Palavrasisoladas
 MMAE:
 MMAL: Morfologica ; sintatica ; semantica ; sintagmasnominais
 MMAM:
 MMRC: vocabulariocontrolado

ID: 48
 AUT: Luiz Claudio Gomes Maia
 TIT: Uso de sintagmas nominais na classificacao automatica de documentos eletronicos
 IPUB:
 ANO: 2008
 FONT: Tese
 INST: UFMG
 CIT: MAIA, L. C. G.
 OOBJ: Aplicaepropoe
 OPRO:
 OAPL: Sistema
 OSMF: OGMA
 MAV: Classificacaoautomatica
 MNC: Textocompleto
 MTC: NoticiaRSS ; trabalhocientifico
 MVAL: Automatica
 MIT: Extracao
 MED: Naoestruturado ; estruturado
 MNLI: Natural
 MNTE: Palavrasisoladas
 MMAE: Stopwords ; pesonumerico

MMAL: Morfológica ; sintática ; semântica ; sintagmas nominais
 MMAM: Cluster
 MMRC:

ID: 49

AUT: Ana Paula Ladeira ; Lidia Alvarenga
 TIT: Processamento de linguagem natural: em busca de evidências
 temáticas nas publicações nacionais e contemporâneas
 IPUB: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO
 ANO: 2009
 FONT: Anais...
 INST: UFMG
 CIT: LADEIRA, A. P. ; ALVARENGA, L.
 OOBJ: Revisão
 OPRO:
 OAPL:
 OSMF:
 MAV:
 MNC:
 MTC:
 MVAL:
 MIT:
 MED:
 MNLI:
 MNTE:
 MMAE:
 MMAL:
 MMAM:
 MMRC:

ID: 50

AUT: Vania Mara Alves Lima ; Vera Regina Casari Boccato
 TIT: O desempenho terminológico dos descritores em ciência da
 informação do vocabulário controlado do SIBI/USP nos processos de
 indexação manual, Automática, e semi-Automática
 IPUB: Perspectivas em Ciência da Informação
 ANO: 2009
 FONT: Periódico
 INST: USP ; UFSCar ; UNESP
 CIT: LIMA, V. M. A. ; BOCCATO, V. R. C.
 OOBJ: Aplicação
 OPRO:
 OAPL: Sistema
 OSMF: SISA
 MAV: indexação manual
 MNC: Resumo

MTC: Trabalhocientifico
 MVAL: Assistida ; semiautomatica ; automatica
 MIT: Extracao
 MED: Naoestruturado
 MNLI: Natural
 MNTE: Palavrasisoladas
 MMAE: Stopwords ; frequencia
 MMAL:
 MMAM:
 MMRC: Vocabulariocontrolado

ID: 51
 AUT: Cristina Miyuki Narukawa ; Isidoro Gil Leiva ; Mariangela Spotti Lopes Fujita
 TIT: Indexacao Automatizada de artigos de periodicos cientificos: analise da aplicacao do software SISA com uso da terminologia DeCS na area de odontologia
 IPUB: Informacao & Sociedade: Estudos
 ANO: 2009
 FONT: Periodico
 INST: UNESP
 CIT: NARUKAWA, C. M. ; GIL LEIVA, I. ; FUJITA, M. S. L.
 OOBJ: Aplica
 OPRO:
 OAPL: Sistema
 OSMF: SISA
 MAV: Indexacaomanual
 MNC: Textocompleto
 MTC: Trabalhocientifico
 MVAL: Semiautomatica
 MIT: Atribuicao
 MED: Naoestruturado
 MNLI: Natural ; controlada
 MNTE: Palavrasisoladas ; termoscompostos
 MMAE: Stopwords ; frequencia
 MMAL:
 MMAM:
 MMRC: Vocabulariocontrolado

ID: 52
 AUT: Graciane Silva Bruzinga Borges
 TIT: Indexacao automatica de documentos textuais: proposta de criterios essenciais
 IPUB:
 ANO: 2009
 FONT: Dissertacao

INST: UFMG
CIT: BORGES, G. S. B.
OOBJ: Revisao
OPRO:
OAPL:
OSMF:
MAV:
MNC:
MTC:
MVAL:
MIT:
MED:
MNLI:
MNTE:
MMAE:
MMAL:
MMAM:
MMRC:

ID: 53
AUT: Viviane Neves dos Santos
TIT: Indexacao automatica de documentos textuais: iniciativas dos grupos
de pesquisa de Universidades Publicas Brasileiras
IPUB:
ANO: 2009
FONT: TCC
INST: USP
CIT: SANTOS, V. N.
OOBJ: Revisao
OPRO:
OAPL:
OSMF:
MAV:
MNC:
MTC:
MVAL:
MIT:
MED:
MNLI:
MNTE:
MMAE:
MMAL:
MMAM:
MMRC:

ID: 54

AUT: Luiz Claudio Gomes Maia ; Renato Rocha Souza
 TIT: Uso de sintagmas nominais na classificacao automatica de documentos eletronicos
 IPUB: Perspectivas em Ciencia da Informacao
 ANO: 2010
 FONT: Periodico
 INST: Faculdade Pitagoras ; UFRJ
 CIT: MAIA, L. C. G. ; SOUZA, R. R.
 OOBJ: Aplicaepropoe
 OPRO:
 OAPL: Sistema
 OSMF: OGMA
 MAV: Classificacaoautomatica
 MNC: Textocompleto
 MTC: NoticiaRSS ; trabalhocientifico
 MVAL: Automatica
 MIT: Extracao
 MED: Naoestruturado ; estruturado
 MNLI: Natural
 MNTE: Palavrasisoladas
 MMAE: Stopwords ; pesonumerico
 MMAL: Morfologica ; sintatica ; semantica ; sintagmasnominais
 MMAM: Cluster
 MMRC:

ID: 55

AUT: Edberto Ferneda ; Maria Cristiane Barbosa Galvao ; Joeli Espirito Santo Rocha
 TIT: Um Metodo de Indexacao automatica de Documentos: aplicacao em laudos de exames radiologicos
 IPUB: ENCONTRO NACIONAL DE PESQUISA EM CIENCIA DA INFORMACAO
 ANO: 2010
 FONT: Anais...
 INST: UNESP ; USP
 CIT: FERNEDA, E. ; GALVAO, M. C. B. ; ROCHA, J. E. S.
 OOBJ: Aplicaepropoe
 OPRO:
 OAPL: Sistema
 OSMF: sintagMed
 MAV: Indexacaomanual
 MNC: Textocompleto
 MTC: Laudo
 MVAL: Automatica
 MIT: Atribuicao
 MED: Naoestruturado
 MNLI: Natural

MNTE: Palavraisoladas ; termoscompostos
 MMAE: Stopwords ; matrizbinaria ; frequencia
 MMAL: Sintagmasnominais
 MMAM:
 MMRC: vocabulariocontrolado

ID: 56
 AUT: Cristina Miyuki Narukawa ; Mariangela Spotti Lopes Fujita
 TIT: O uso de linguagens documentarias alfabeticas na indexacao automatizada
 IPUB: SEMINARIO NACIONAL DE BIBLIOTECAS UNIVERSITARIAS
 ANO: 2010
 FONT: Anais...
 INST: UNESP ; USP
 CIT: NARUKAWA, C. M. ; FUJITA, M. S. L.
 OOBJ: Propoe
 OPRO: Sistema
 OAPL:
 OSMF: SISA
 MAV: Indexacaomanual
 MNC: Textocompleto
 MTC: Trabalhocientifico
 MVAL: Semiautomatica
 MIT: Atribuicao
 MED: Naoestruturado
 MNLI: Natural ; controlada
 MNTE: Palavraisoladas ; termoscompostos
 MMAE: Stopwords ; frequencia
 MMAL:
 MMAM:
 MMRC: Vocabulariocontrolado

ID: 57
 AUT: Graciane Silva Bruzuinga Borges
 TIT: Indexacao automatica de documentos textuais: proposta de criterios essenciais
 IPUB: ENCONTRO NACIONAL DE PESQUISA EM CIENCIA DA INFORMACAO
 ANO: 2010
 FONT: Anais...
 INST: UFMG
 CIT: BORGES, G. S. B.
 OOBJ: Revisao
 OPRO:
 OAPL:
 OSMF:
 MAV:

MNC:
MTC:
MVAL:
MIT:
MED:
MNLI:
MNTE:
MMAE:
MMAL:
MMAM:
MMRC:

ID: 58

AUT: Ana Paula Ladeira

TIT: Processamento de linguagem natural: caracterizacao da producao
cientifica dos pesquisadores brasileiros

IPUB:

ANO: 2010

FONT: Tese

INST: UFMG

CIT: LADEIRA, A. P.

OOBJ: Revisao

OPRO:

OAPL:

OSMF:

MAV:

MNC:

MTC:

MVAL:

MIT:

MED:

MNLI:

MNTE:

MMAE:

MMAL:

MMAM:

MMRC:

ID: 59

AUT: Renato Fernandes Correa ; Darliane Goes de Miranda ; Camila
Oliveira de Almeida Lima ; Tiago Jose da Silva

TIT: Indexacao e recuperacao de teses e dissertacoes por meio de
sintagmas nominais

IPUB: AtoZ

ANO: 2011

FONT: Periodico

INST: UFPE
 CIT: CORREA, R. F. ; MIRANDA, D. G. ; LIMA, C. O. A. ; SILVA, T. J.
 OOBJ: Aplicaepropoe
 OPRO:
 OAPL: Sistema
 OSMF: OGMA
 MAV: Indexacaomanual
 MNC: Resumo
 MTC: Trabalhocientifico
 MVAL: Automatica
 MIT: Extracao
 MED: Naoestruturado
 MNLI: natural
 MNTE: Palavraisoladas ; termoscompostos
 MMAE:
 MMAL: Sintagmasnominais
 MMAM:
 MMRC:

ID: 60
 AUT: Cristina Miyuki Narukawa
 TIT: Estudo de vocabulario controlado na indexacao automatica: aplicacao no processo de indexacao do Sistema de Indizaco Semiautomatica (SISA)
 IPUB:
 ANO: 2011
 FONT: Dissertacao
 INST: UNESP
 CIT: NARUKAWA, C. M.
 OOBJ: Aplicaepropoe
 OPRO:
 OAPL: Sistema
 OSMF: SISA
 MAV: Indexacaomanual
 MNC: Textocompleto
 MTC: Trabalhocientifico
 MVAL: Semiautomatica
 MIT: Atribuicao
 MED: Naoestruturado
 MNLI: Natural ; controlada
 MNTE: Palavraisoladas ; termoscompostos
 MMAE: Stopwords ; frequencia
 MMAL:
 MMAM:
 MMRC: Vocabulariocontrolado

ID: 61

AUT: Ardala Ponce Kochani ; Vera Regina Casari Boccato ; Milena Polsinelli Rubi
TIT: Politica de indexacao para sistemas automatizados de coordenadorias de comunicacao social em ambientes universitarios
IPUB: Congresso Brasileiro de Biblioteconomia, Documentacao e Ciencia da Informacao
ANO: 2011
FONT: Anais...
INST: UFSCar
CIT: KOCHANI, A. P. ; BOCCATO, V. R. C. ; RUBI, M. P.
OOBJ: Revisao
OPRO:
OAPL:
OSMF:
MAV:
MNC:
MTC:
MVAL:
MIT:
MED:
MNLI:
MNTE:
MMAE:
MMAL:
MMAM:
MMRC:

ID: 62
AUT: Cinthia Holanda ; Marcia Ivo Braz
TIT: Indexacao automatica de conteudos na web: analise de sites de museus
IPUB: Biblionline,
ANO: 2012
FONT: Periodico
INST: UFPE
CIT: HOLANDA, C. ; BRAZ, M. I.
OOBJ: Revisao
OPRO:
OAPL:
OSMF:
MAV:
MNC:
MTC:
MVAL:
MIT:
MED:

MNLI:
MNTE:
MMAE:
MMAL:
MMAM:
MMRC:

ID: 63
AUT: Edberto Fernalda
TIT: Introducao aos modelos computacionais de recuperacao de informacao
IPUB: Editora Ciencia Moderna
ANO: 2012
FONT: Livro
INST: UNESP
CIT: FERNEDA, E.
OOBJ: Revisao
OPRO:
OAPL:
OSMF:
MAV:
MNC:
MTC:
MVAL:
MIT:
MED:
MNLI:
MNTE:
MMAE:
MMAL:
MMAM:
MMRC:

ID: 64
AUT: Frederick Wilfrid Lancaster
TIT: Indexacao e resumo: teoria e pratica
IPUB: Briquet de Lemos / Livros
ANO: 2004
FONT: Livro
INST: Universidade de Illinois
CIT: LANCASTER, F. W.
OOBJ: Revisao
OPRO:
OAPL:
OSMF:
MAV:

MNC:
MTC:
MVAL:
MIT:
MED:
MNLI:
MNTE:
MMAE:
MMAL:
MMAM:
MMRC:

ID: 65
AUT: Jennifer Rowley
TIT: A biblioteca eletronica
IPUB: Briquet de Lemos / Livros
ANO: 2002
FONT: Livro
INST: School of Management and Social Sciences
CIT: ROWLEY, J.
OOBJ: Revisao
OPRO:
OAPL:
OSMF:
MAV:
MNC:
MTC:
MVAL:
MIT:
MED:
MNLI:
MNTE:
MMAE:
MMAL:
MMAM:
MMRC:

ID: 66
AUT: Claire Guinchat ; Michel Menou
TIT: Introducao geral ...s ciencias e tecnicas da informacao e documentacao
IPUB: IBICT
ANO: 1994
FONT: Livro
INST: UNESCO
CIT: GUINCHAT, C. ; MENU, M.
OOBJ: Revisao

OPRO:
 OAPL:
 OSMF:
 MAV:
 MNC:
 MTC:
 MVAL:
 MIT:
 MED:
 MNLI:
 MNTE:
 MMAE:
 MMAL:
 MMAM:
 MMRC:

ID: 67
 AUT: Mariangela Spotti Lopes Fujita
 TIT: Precis na lingua portuguesa: teoria e pratica de Indexacao
 IPUB: Editora Universidade de Brasilia
 ANO: 1989
 FONT: Livro
 INST: UNB
 CIT: FUJITA, M. S. L.
 OOBJ: Aplica
 OPRO:
 OAPL: Sistema
 OSMF: PRECIS
 MAV: Indexacaomanual
 MNC: Textocompleto
 MTC: Noticia ; trabalhocientifico
 MVAL: Semiautomatica
 MIT: Atribuicao
 MED: Naoestruturado
 MNLI: Natural
 MNTE: Palavrasisoladas ; termoscompostos
 MMAE: Frequencia
 MMAL: Sintatica ; semantica
 MMAM:
 MMRC: Vocabulariocontrolado

ID: 68
 AUT: Ligia Cafe
 TIT: Contribuicoes da gramatica funcional da delimitacao de segmentos
 descritores de informacao
 IPUB: Thesaurus

ANO: 2003
 FONT: Capitulo de Livro
 INST: UFSC
 CIT: CAFE, L.
 OOBJ: Propoe
 OPRO: Metodo
 OAPL:
 OSMF: UTC
 MAV:
 MNC: Textocompleto
 MTC: Trabalhocientifico
 MVAL: Semiautomatica
 MIT: Extracao
 MED: Naoestruturado
 MNLI: Natural
 MNTE: Palavrasisoladas ; termoscompostos
 MMAE:
 MMAL: Morfologica ; sintatica ; semantica ; sintagmasnominais
 MMAM:
 MMRC:

ID: 69
 AUT: Claudio Gottschalg-Duque
 TIT: SiRILiCO Uma Proposta para um Sistema de Recuperacao de Informacao baseado em Teorias da Linguistica Computacional e Ontologia
 IPUB:
 ANO: 2005
 FONT: Tese
 INST: UFMG
 CIT: GOTTSCHALG-DUQUE, C
 OOBJ: Aplicaepropoe
 OPRO:
 OAPL: Sistema
 OSMF: SiRILiCO
 MAV: participacaodousuario
 MNC: Textocompleto
 MTC: Trabalhocientifico
 MVAL: Automatica
 MIT: Extracao
 MED: Textonaoestruturado
 MNLI: Natural
 MNTE: Palavrasisoladas
 MMAE:
 MMAL: Morfologica ; sintatica ; semantica ; sintagmasnominais
 MMAM:
 MMRC: Ontologia

APÊNDICE C – Ficha de análise de conteúdo

Categoria Objetivo			
	Proposição ☐ Revisão de literatura ☐ Sistema ☐ Método ☐ Fórmula	Aplicação ☐ Sistema ☐ Método ☐ Fórmula	Aplicação e proposição ☐ Sistema ☐ Método ☐ Fórmula
Nome do Sistema/Método/Fórmula			
Categoria Aspectos Metodológicos			
Avaliação: ☐ Comparação com a indexação manual ☐ Participação do usuário ☐ Classificação automática ☐ Julgamento do autor			
Natureza ☐ Resumo ☐ Citações ☐ Título ☐ Verbetes ☐ Texto completo			
Tipologia ☐ Carta ☐ Trabalho científico ☐ Manuscrito ☐ Ata ☐ Notícia ☐ Documento oficial ☐ Laudo ☐ Livro ☐ Nota técnica ☐ Nota técnica ☐ Notícia RSS ☐ Peça teatral			
Análise dos Sistemas/Métodos/Fórmulas de Indexação Automática			
Validação dos termos ☐ Assistida por computador ☐ Semi-automática ☐ Automática	Identificação dos termos ☐ Extração ☐ Atribuição	Entrada de dados ☐ Texto não-estruturado ☐ Texto estruturado ☐ Texto com marcas	
Linguagem de indexação			
Natureza da linguagem ☐ Natural ☐ Controlada Natureza dos termos ☐ Palavras isoladas ☐ Termos compostos			
Identificação/Ponderação/Seleção de termos			
Análise estatística ☐ Radicalização ☐ Eliminação de stopwords ☐ Peso numérico ☐ Matriz binária ☐ Análise de co-ocorrência ☐ Análise de frequência ☐ Localização ☐ Dicionário de raízes	Análise linguística ☐ Análise sintática ☐ Análise morfológica ☐ Análise semântica ☐ Sintagmas nominais	Aprendizado de máquina ☐ Classificação ☐ Análise de cluster	Representação do conhecimento ☐ Vocabulário controlado ☐ Ontologia ☐ Gráfico conceitual