UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE ARTES E COMUNICAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

MÁRCIO HENRIQUE WANDERLEY FERREIRA

PERCURSO METODOLÓGICO PARA A FORMULAÇÃO DE INDICADORES TEMÁTICOS DE INFORMAÇÃO CIENTÍFICA: APORTES DA INDEXAÇÃO AUTOMÁTICA E DOS ESTUDOS MÉTRICOS DA INFORMAÇÃO

RECIFE
2021

MÁRCIO HENRIQUE WANDERLEY FERREIRA

PERCURSO METODOLÓGICO PARA A FORMULAÇÃO DE INDICADORES TEMÁTICOS DE INFORMAÇÃO CIENTÍFICA: APORTES DA INDEXAÇÃO AUTOMÁTICA E DOS ESTUDOS MÉTRICOS DA INFORMAÇÃO

Tese de doutorado apresentada ao Programa de Pós-graduação em Ciência da Informação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de doutor em Ciência da Informação.

Área de concentração: Informação, Memória e Tecnologia.

Orientador: Prof. Dr. Renato Fernandes Corrêa.

RECIFE

2021

Catalogação na fonte
Bibliotecária Jéssica Pereira de Oliveira – CRB-4/2223

F383p Ferreira, Márcio Henrique Wanderley
Percurso metodológico para a formulação de indicadores temáticos de informação científica: aportes da Indexação Automática e dos Estudos Métricos da Informação / Márcio Henrique Wanderley Ferreira. – Recife, 2021.
256f.: il., tab.

Sob orientação de Renato Fernandes Corrêa.
Tese (Doutorado) – Universidade Federal de Pernambuco. Centro de Artes e Comunicação. Programa de Pós-Graduação em Ciência da Informação, 2021.

Inclui referências, apêndices e anexos.

1. Estudos Métricos da Informação. 2. Indexação Automática. 3. Mineração de texto. 4. Percurso metodológico. 5. Indicadores temáticos. I. Corrêa, Renato Fernandes (Orientação). II. Título.

020 CDD (22. ed.) UFPE (CAC 2022-04)

MÁRCIO HENRIQUE WANDERLEY FERREIRA

PERCURSO METODOLÓGICO PARA A FORMULAÇÃO DE INDICADORES TEMÁTICOS DE INFORMAÇÃO CIENTÍFICA: APORTES DA INDEXAÇÃO AUTOMÁTICA E DOS ESTUDOS MÉTRICOS DA INFORMAÇÃO

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

Aprovada em: 27/09/2021

BANCA EXAMINADORA

Prof. Dr. Renato Fernandes Corrêa (Orientador)

Universidade Federal de Pernambuco

Prof. Dr. Fábio Mascarenhas e Silva (Examinador Interno)

Universidade Federal de Pernambuco

Prof. Dr. Murilo Artur Araújo da Silveira (Examinador Interno)

Universidade Federal de Pernambuco

Prof. Dr. Isidoro Gil Leiva (Examinador Externo)

Universidade de Murcia

Prof. Dr. Fábio Castro Gouveia (Examinador Externo)

Fiocruz-RJ

Dedico à pacificação da humanidade, por uma união universal e equilíbrio entre as civilizações.

AGRADECIMENTOS

À fonte criadora universal e às leis fundamentais que sustentam a vida no Universo.

À **UFPE** e ao **PPGCI** por todo apoio institucional, estrutura e corpo técnico que sempre estiveram à disposição.

À instituição de fomento **CAPES**, que forneceu o apoio por meio da bolsa de doutorado durante 2 anos e 6 meses.

Ao meu Orientador, **Renato Fernandes Corrêa**, por toda parceria e apoio em momentos cruciais.

Ao orientando do professor Renato, **Ramon**, por ter auxiliado na metodologia desta pesquisa.

Aos professores **Fábio Mascarenhas**, **Murilo Silveira**, **Fábio Gouveia** e **Isidoro Gil Leiva** por terem aceitado participar desta banca de defesa.

Aos incríveis amigos que fiz na Universidade, desde que entrei na UFPE em agosto de 2009.

À minha prima querida, **Letícia Quintella**, por ter auxiliado em momentos crucias.

À minha amiga, **Tatyane Lúcia Cruz** por toda amizade e carinho.

Ao meu amigo, **Natanael Vitor Sobral** por todo o apoio e estímulo.

Ao meu pai, **Márcio Augusto Ferreira**, e à minha mãe, *in memoriam*, **Eufrásia Wanderley**, por terem me fornecido todo o apoio incondicional para que eu chegasse a este momento.

À minha madrinha, **Vera Lúcia de Freitas Wanderley**, por todo apoio incondicional na minha trajetória.

À minha companheira, **Críssia Santana**, por ter me apoiado em todos os instantes.

E a todos os meus entes queridos e familiares que contribuíram direta ou indiretamente para que fosse concretizada essa jornada.

FERREIRA, Márcio Henrique Wanderley. **Percurso Metodológico para a formulação de Indicadores Temáticos de Informação Científica**: Aportes da Indexação Automática e dos Estudos Métricos da Informação. Recife, 2021. Tese (Doutorado em Ciência da Informação) – Programa de Pós-graduação em Ciência da Informação, Centro de Artes e Comunicação, Universidade Federal de Pernambuco, Recife, 2021, 257f.

RESUMO

Esta tese objetiva propor um modelo metodológico para a formulação de indicadores temáticos de informação científica, a partir da integração teórica e metodológica entre a Indexação Automática e os Estudos Métricos da Informação. Realiza uma pesquisa metodológica de natureza aplicada, utilizando procedimentos técnicos dos Estudos Métricos da Informação e da Indexação Automática. Utiliza, da aplicação do sistema maui de indexação automática, como instrumento para a categorização das palavras-chave dos registros bibliográficos em conceitos de um tesauro de especialidade, contemplando, nos campos de metadados dos registros, termos no idioma do texto informado, e, como um dos instrumentos escolhidos para aplicação da metodologia. Adota, o software microsoft excel como ferramenta para elaboração dos gráficos de frequência de palavras. Em seguida, emprega o software Iramuteq, para aplicar os Indicadores Temáticos nos processos de análise, nos resultados obtidos pela Indexação automática. Pretende validar o percurso metodológico proposto em dois *corpora*, a saber: dois conjuntos de registros bibliográficos de artigos de periódicos brasileiros. O primeiro contempla um conjunto de 60 artigos utilizados por Souza (2005); o segundo, 82 artigos sobre o assunto "Biblioteca Digital" utilizados por Ferreira e Corrêa (2018). Analisa a generalização e a extrapolação do percurso metodológico proposto para a realização do estudo métrico das temáticas em outras áreas da ciência brasileira. Os principais resultados apontam para as métricas de indexação obtidas em ambos os *corpora*, após a adoção do maui. Indica as estatísticas de ocorrência e frequência, observando a dispersão dos termos nos documentos. Evidencia os gráficos de indicadores temáticos, observando o comportamento terminológico da frequência dos termos e coocorrência dentro dos *corpora* observados. E finalmente, defende a adoção do percurso metodológico proposto considerando todas as etapas envolvidas.

Palavras-chave: estudos métricos da informação; indexação automática; mineração de texto; percurso metodológico; indicadores temáticos.

FERREIRA, Márcio Henrique Wanderley. **Methodological path for the formulation of Thematic Indicators of Scientific Information**: Support for Automatic Indexing and Metric Studies of Information. Recife, 2021. Thesis (Doctorate in Information Science) - Postgraduate Program in Information Science, Arts and Communication Center, Federal University of Pernambuco, Recife, 2021, 257f.

ABSTRACT

This thesis aims to propose a methodological model for the formulation of thematic indicators of scientific information, based on the theoretical and methodological integration between Automatic Indexing and Metric Information Studies. It conducts an exploratory research of an applied nature, using technical procedures from the studies of Information Metrics and Automatic Indexing. It uses, from the application of the maui automatic indexing system, as an instrument for categorizing the keywords of bibliographic records in concepts of a specialty thesaurus, including, in the records' metadata fields, terms in the language of the informed text, and, as one of the instruments chosen to apply the methodology. Uses microsoft excel software as a tool for drawing up word frequency graphs. Then, it adopts the Iramuteq software, to apply the Thematic Indicators in the analysis processes, given as input the results obtained by automatic indexing process. It intends to validate the proposed methodological pathway in two *corpora*, namely two sets of bibliographic records of articles from Brazilian journals. The first includes a set of 60 articles used by Souza (2005); the second, 82 articles on the subject "Digital Library" used by Ferreira and Corrêa (2018). Finally, this work analyzes the generalization and extrapolation of the proposed methodological pathway for conducting the metric study of the themes in other areas of Brazilian science. The main results point to the indexing metrics obtained in both corpora, after the adoption of maui. Indicates occurrence and frequency statistics, noting the dispersion of terms in the documents. Point out the thematic indicator graphs, observing the terminological behavior of the frequency of terms and co-occurrence within the observed corpora. And finally, it defends the adoption of the proposed methodological path considering all the steps involved.

Keywords: information metrics; automatic indexing; text mining; methodological pathway; thematic indicators.

LISTA DE ILUSTRAÇÕES

Quadro 1 –	Principais Leis e Estudos da Bibliometria.....	28
Quadro 2 –	Definições das 3 principais disciplinas inseridas nos EMI.....	30
Quadro 3 –	Finalidades e objetos das disciplinas inseridas nos EMI.....	31
Figura 1 –	Divisões entre as disciplinas no campo dos EMI.....	32
Quadro 4 –	Áreas de concentração da Bibliometria e Cienciometria.....	33
Quadro 5 –	Indicadores e utilidades para a análise da produção científica.....	36
Figura 2 –	Descrição da Busca 1.....	47
Figura 3 –	Descrição da Busca 2.....	48
Quadro 6 –	19 trabalhos relevantes identificados.....	54
Figura 4 –	Etapas da Mineração de Textos.....	66
Quadro 7 –	Tarefa de mineração de texto.....	67
Figura 5 –	Processo Cognitivo da Indexação.....	73
Quadro 8 –	Tipos de Indexação Automática.....	83
Gráfico 1 –	Tendência de produções de trabalhos ao longo de 25 anos.....	84
Gráfico 2 –	Quantidade de trabalhos dos 25 autores mais produtivos.....	85
Gráfico 3 –	Índice de Citações dos 20 artigos mais representativos.....	87
Gráfico 4 –	Lista das 25 áreas do conhecimento com maior número de publicações.....	89
Gráfico 5 –	Quantidade de trabalhos dos 20 autores mais produtivos.....	90
Quadro 9 –	Comparação entre os principais softwares de extração automática de termos.....	92
Quadro 10 –	Técnicas de detecção de palavras-chave descritas no artigo “Keyword Detection Techniques: A Comprehensive Study”.....	100
Figura 6 –	Fluxograma da Indexação Automática.....	107
Figura 7 –	Fluxograma do Estudo Métrico.....	108
Quadro 11 –	Pressupostos a serem verificados na validação do percurso.....	115
Quadro 12 –	Objetivos Específicos Verificados na Pesquisa.....	116
Figura 8 –	Arquivos Formatados para Análise.....	118
Figura 9 –	Fluxo de Etapas Adotadas na Indexação Intelectual.....	118
Figura 10 –	Exemplo de Arquivo com extensão .key formatado.....	119
Figura 11–	Exemplo de arquivo com extensão .txt formatado.....	119

Gráfico 6 – Estatística do *Corpus* A com os Metadados (Título, Resumo e Palavras-Chave do Autor)..... 134
Gráfico 7 – 20 palavras mais frequentes do *Corpus* A..... 135
Gráfico 8 – Análise estatística do *Corpus* B..... 136
Gráfico 9 – 20 Palavras mais frequentes do *Corpus* B..... 137
Figura 12 – Análise de Agrupamentos (Clusters) do *Corpus* A com os Metadados (Título, Resumo e Palavras-Chave do Autor)..... 140
Gráfico 10 – Palavras mais frequentes do *Corpus* A (Indexação do Autor)..... 142
Figura 13 – Análise de Agrupamentos (Clusters) do *Corpus* A com os Metadados (Título, Resumo e Palavras-Chave do Maui)..... 143
Gráfico 11 – Palavras mais frequentes do *Corpus* A (Indexação do Maui)..... 146
Gráfico 12 – Estatística dos Indexadores Manuais – *Corpus* A..... 147
Gráfico 13 – 22 descritores mais frequentes (palavras isoladas) - Indexadores Manuais – *Corpus* A..... 148
Figura 14 – Análise de Agrupamentos (Clusters) do *Corpus* A com os Metadados (Título, Resumo e Palavras-Chave do Indexador Manual)..... 149
Gráfico 14 – 25 descritores mais frequentes (palavras isoladas/compostas) - Indexadores Manuais – *Corpus* A..... 151
Gráfico 15 – Estatística dos Indexadores Manuais – *Corpus* B..... 152
Gráfico 16 – 20 palavras mais frequentes (palavras isoladas) - Indexadores Manuais – *Corpus* B..... 153
Gráfico 17 – 24 descritores mais frequentes (palavras isoladas/compostas) - Indexadores Manuais – *Corpus* B..... 153
Figura 15 – Análise de Agrupamentos (Clusters) do *Corpus* B com os Metadados (Título, Resumo e Palavras-Chave do Indexador Manual)..... 154
Figura 16 – Análise de Agrupamentos (Clusters) do *Corpus* B com os Metadados (Título, Resumo e Palavras-Chave do Autor)..... 156
Figura 17 – Análise de Agrupamentos (Clusters) do *Corpus* B com os Metadados (Título, Resumo e Palavras-Chave do Maui)..... 158
Gráfico 18 – Frequência de Palavras-chave atribuídas manualmente do *Corpus* A dos Autores..... 159

Gráfico 19 – Frequência de descritores do Maui para o *Corpus* A..... 160
Gráfico 20 – Frequência de Palavras-chave manuais do *Corpus* B..... 161
Gráfico 21- Frequência de Palavras-chave do Maui do *Corpus* B..... 162

LISTA DE TABELAS

Tabela 1 – Autores dos 20 Artigos Mais Citados..... 88
Tabela 2 – Resultados do Maui no *Corpus* A..... 121
Tabela 3 – Média das Métricas Obtidas pelo Maui no *Corpus* A..... 124
Tabela 4 – Resultados Extraídos do Maui no *Corpus* B..... 127
Tabela 5 – Média das Métricas Obtidas pelo Maui no *Corpus* B..... 132
Tabela 6 – Análise Comparativa entre os 7 termos manuais mais frequentes do *Corpus* A..... 160
Tabela 7 – Análise Comparativa entre os 7 termos manuais mais frequentes do *Corpus* B..... 163

LISTA DE SIGLAS

CAPES	Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior
CNEN	Comissão Nacional de Energia Nuclear
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CT&I	Ciência Tecnologia e Inovação
DC	Descoberta de Conhecimento
EMI	Estudos Métricos da Informação
ENANCIB	Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação
IA	Indexação Automática
IM	Indexação Manual
IT	Indicadores Temáticos
KWIC	Keyword in Context
MT	Mineração de Texto
RI	Recuperação da Informação
SRI	Sistemas de Recuperação da Informação
TBCI	Tesauro Brasileiro em Ciência da Informação
TIC's	Tecnologias da Informação e da Comunicação
WoS	*Web of Science*

SUMÁRIO

1	**INTRODUÇÃO**	**15**
2	**FUNDAMENTAÇÃO TEÓRICA**	**25**
2.1	**Estudos Métricos da Informação**	**25**
2.1.1	Análise temática e suas relações	34
2.1.2	Trabalhos Relacionados a Indicadores Temáticos	40
2.1.2.1	*Cartografia Temática e Cartografia Bibliométrica*	40
2.1.2.2	*Bibliometria Temática e Mapeamento Bibliométrico*	43
2.1.3	Indexação Automática aplicada ao Mapeamento Bibliométrico	47
2.1.4	Definição sobre o software de Estudos Métricos – Iramuteq	61
2.2	**Descoberta de Conhecimento em Texto**	**62**
2.2.1	Estudos Sobre Mineração de Textos	67
2.3	**Indexação Automática**	**69**
2.3.1	Observação do Estado da Arte	83
2.3.2	Identificação dos Sistemas de Indexação Automática ou de Extração de Palavras-chave	91
2.3.3	Definição sobre o software de Indexação Automática – Maui	102
3	**PROCEDIMENTOS METODOLÓGICOS**	**105**
3.1	**Caracterização da pesquisa**	**105**
3.2	**Percurso Metodológico Proposto**	**106**
3.2.1	Descrição das etapas do processo de proposição do percurso metodológico	106
3.3	**Compilação dos Corpora**	**110**
3.4	**Procedimentos necessários para a compilação de dados**	**111**
3.4.1	Análise de Dados por meio da Ferramenta de Indexação Automática – Maui	111
3.4.2	Treinamento e Execução do Software Maui	112
3.5	**Análise de Dados por meio da Ferramenta dos Estudos Métricos – Iramuteq**	**113**
3.5.1	Descrição do funcionamento e Configuração do Software Iramuteq	113
3.6	**Análise do percurso metodológico**	**114**
4	**ANÁLISE DOS RESULTADOS**	**117**

4.1 Resultados da Indexação Automática..... 117
4.2 Análises Estatísticas dos Corpora..... 134
4.3 Resultados obtidos do Fluxograma do Estudo Métrico..... 137
5 CONCLUSÃO..... 164
REFERÊNCIAS..... 168
APÊNDICE A – RESULTADOS DO MAUI DO *CORPUS* A..... 185
APÊNDICE B – RESULTADOS DO MAUI DO *CORPUS* B..... 196
APÊNDICE C – LISTA DE FREQUÊNCIA DE TÓPICOS INDEXADOS PELO MAUI NOS CORPORA..... 212
APÊNDICE D – COMPARAÇÃO ENTRE AS PALAVRAS-CHAVE E OS TÓPICOS INDEXADOS – *CORPUS* A..... 218
APÊNDICE E – COMPARAÇÃO ENTRE AS PALAVRAS-CHAVE E OS TÓPICOS INDEXADOS – *CORPUS* B..... 230
ANEXO A – TABELA COM AS REFERÊNCIAS DOS 60 ARTIGOS DO *CORPUS* A-SOUZA..... 245
ANEXO B – TABELA COM AS REFERÊNCIAS DOS 82 ARTIGOS DO *CORPUS* B – FERREIRA E CORRÊA..... 250

1 INTRODUÇÃO

Entre cientistas e formuladores de políticas em Ciência Tecnologia e Inovação (CT&I), é aceito, de forma expressiva, que a pesquisa básica e a pesquisa aplicada constituem-se como forças fundamentais capazes de propiciar o desenvolvimento científico e tecnológico de uma nação. Contudo, a promoção dessas pesquisas necessita de recursos adequados para análise e monitoramento, sobretudo, da produção em CT&I.

Tendo isso em vista, nota-se que os Estudos Métricos da Informação (EMI) surgem como um domínio do conhecimento que mensura e dimensiona as informações científicas produzidas. Por sua vez, esse domínio serve de insumo para o estabelecimento de diretrizes em políticas de CT&I, contribuindo, inclusive, para a definição das áreas que devem receber preferência quando da destinação dos recursos.

Assim, o desenvolvimento de estudos científicos por meio dos EMI encontra na Ciência da Informação (CI) instrumentos apropriados. De acordo com Le Coadic (2004), a CI é uma ciência interdisciplinar que compreende a colaboração de diversos campos do conhecimento, a saber: Sociologia, Informática, Matemática, Lógica, Estatística, Economia, Linguística e outros. Nesse sentido, particularmente na disciplina específica dos EMI, a CI torna-se um ramo científico apto a estudar os desdobramentos que as análises métricas proporcionam.

Para Oliveira e Grácio (2011, p. 19, grifo do autor):

> os “Estudos Métricos” compreendem o conjunto de estudos relacionados à avaliação da informação produzida, mais especialmente científica, em diferentes suportes, baseados em recursos quantitativos como ferramentas de análise. Fundamentados na sociologia da ciência, na ciência da informação, matemática, estatística e computação, são estudos de natureza teórico-conceitual, quando contribuem para o avanço do conhecimento da própria temática, propondo novos conceitos e indicadores, bem como reflexões e análises relativas à área. São, também, de natureza metodológica, quando se propõem a dar sustentação aos trabalhos de caráter teórico da área onde estão aplicados.

Isto posto, assume-se que os EMI possibilitam oportunidades descritivas sobre o desempenho de determinadas áreas do conhecimento, principalmente quando combinados a outros domínios das Tecnologias da Informação e Comunicação (TICs).

Oliveira e Grácio (2011, p. 19) afirmam que os EMI surgem da união de disciplinas como: Bibliometria, Cientometria, Webometria e Infometria, sendo a Altmetria mais recentemente acrescentada. Os EMI congregam instrumentos capazes de identificar realidades antes desconhecidas, provocando reflexões sobre comportamentos científicos individuais e

coletivos visíveis apenas pela aplicação das métricas informacionais. Tais instrumentos podem favorecer o progresso do conhecimento da CI, indicando conjunturas e cenários apenas conhecidos pela aplicação de indicadores científicos previamente determinados.

Todavia, apesar da capacidade extensiva dos EMI, eles necessitam de domínios do conhecimento que possam facilitar a compreensão sobre determinado campo. Os estudos sobre os Indicadores Temáticos (IT) são um caminho válido para contribuir no processo de entendimento sobre um determinado *corpus* bibliográfico, uma vez que possuem a capacidade de representar, em linhas gerais, o conhecimento sobre os temas tratados em documentos ou áreas desconhecidas por meio de visualizações gráficas. Kobashi e Santos (2006, p.31) tratam de IT quando afirmam que, "a utilidade da visualização de dados por meio de mapas tem respaldo em estudos sobre a percepção, que mostram que o ser humano tem primeiro uma percepção global de uma cena antes de atentar para os detalhes". Nesse contexto, o ser humano identifica de forma mais compreensível uma visualização gráfica em mapas ou gráficos, tendo seu entendimento facilitado sobre assuntos considerados complexos.

Observando os aspectos da evolução do armazenamento de grande quantidade de dados e o fenômeno *big data*, pode-se inferir que as técnicas de visualização da informação podem auxiliar na análise de padrões observados nos dados, mais especificamente, nesta pesquisa, o de padrões temáticos de comportamento das palavras-chave. Tais técnicas foram mencionadas por Freitas *et al*. (2001, p. 144), quando afirmam:

> Usuários acessando essas grandes e diversificadas bases de dados ou realizando buscas na internet obtêm facilmente um volume enorme de informações, dentre as quais muitas podem ser irrelevantes para os objetivos da tarefa sendo realizada. Assim, a sobrecarga de informações é uma das principais preocupações na representação de resultados obtidos através de mecanismos de recuperação de informações. Uma abordagem para contornar as dificuldades de selecionar as informações relevantes dentre os resultados de busca é utilizar técnicas de visualização de informações através das quais o usuário obtém uma representação visual que, se por um lado abstrai detalhes do conjunto de informações, por outro propicia uma organização desse conjunto segundo algum critério.

Assim, com o objetivo de oferecer uma experiência usual e acessível aos usuários, os sistemas de visualização de informações fornecem graficamente dados que podem ser percebidos e compreendidos, neste caso, a análise e interpretação dos resultados torna-se possível. Kobashi e Santos (2006, p.32), afirmam que os indicadores podem ser definidos como: "dados estatísticos que representam aspectos da realidade". Esses indicadores, inicialmente, poderiam ser divididos em indicadores de produção (número de publicações por tipo de documento); indicadores de citação (contagem de citações recebidas por um artigo em

um periódico); indicadores de ligação (coocorrência de autoria, citações e palavras). Nesta pesquisa, pretendeu-se utilizar a noção de indicadores de ligação com a coocorrência de palavras para a construção dos IT, portanto, a definição de Kobashi e Santos (2006), aproxima-se da definição conceitual sobre IT adotada neste documento que é a visualização temática da representação conceitual sobre um determinado conjunto de termos.

Nessa conjuntura, a atribuição de IT, surge como uma via possível, não apenas pela possibilidade de tornar-se uma técnica eficiente, mas, também, por beneficiar a visualização da informação. Contudo, definir qual termo representa de forma mais precisa determinado conjunto de documentos não é uma tarefa simples, exigindo do pesquisador conhecimentos de classificação e indexação. Ferneda (2003) descreve essa realidade, quando apresenta a dificuldade existente na descrição do conteúdo de um documento. O autor afirma que apesar de existirem sistemas computacionais que facilitem o potencial representativo de um termo indexado, existe a dificuldade em atribuir bons termos de indexação. Kobashi, Diaz e Santana (2014, p. 38), na mesma linha, chamam a atenção para a ausência de dados estruturados nas bases como um fator limitante e que pode dificultar o acesso a determinadas tipologias de informação.

Dessa forma, a indexação apresenta-se como uma técnica capaz de atribuir conceitos (indexadores) a um determinado grupo de documentos. Lancaster (2004), afirmou em seu livro que o principal propósito da indexação é o de construir uma lista de representações dos documentos publicados, esses documentos podem estar agrupados numa base de dados ou dispersos em diversas fontes de informação. No entanto, indexar manualmente pode ser uma tarefa dispendiosa e exaustiva, pois dependendo do volume de textos a serem lidos e compreendidos, o processo pode se tornar desafiador. Por esse motivo a Indexação Automática (IA) torna-se o aprimoramento da Indexação manual, e seu objetivo é tornar o processo mais rápido e menos custoso para todos os indivíduos envolvidos.

Nessa lógica, as técnicas da Indexação Automática procuram agilizar o processo de descoberta de termos relevantes, tornando mais eficiente o planejamento de estratégias de busca que envolvem ampla complexidade, inclusive na formulação de metodologias. Esses métodos viabilizam a pesquisa por termos de linguagem natural nos títulos, resumos e textos completos, bem como na procura por termos de linguagem controlada nos descritores, anos de publicação e autores. Contudo, tais procedimentos esbarram na dificuldade de extrair conceitos dos diversos blocos textuais devido à dificuldade de identificar esses termos relevantes. Por isso, tal limitação torna a indexação intelectual um instrumento importante, por atribuir termos chave relacionados à semântica textual (LOPES, 2002).

Diante da problemática supramencionada e das potencialidades apresentadas pela IA em conjunto com os EMI para o alcance de soluções, esta pesquisa parte da seguinte pergunta norteadora: **como se configuram as relações entre a IA e os EMI para a formulação metodológica de indicadores temáticos de informação científica?** Ressalta-se que neste bojo contemplam-se as questões relacionadas ao progresso do potencial das técnicas de IA amparando-se em tesauros e na estruturação de *corpus* documentais legíveis por máquina que favorecem os processos automatizados de indexação.

Visando responder ao questionamento macro apontado, realiza-se, num primeiro momento, uma criteriosa revisão de literatura nas principais bases de dados internacionais, buscando os principais *softwares* livres utilizados no processo de automatização da indexação. Importou não apenas identificar esses *softwares*, mas, também, analisá-los e escolher os que melhor se adaptavam à proposta desta pesquisa. Em seguida, fez-se necessário investigar quais *softwares*/técnicas procedentes dos EMI são aptos para apresentar os dados de forma compreensível e fidedigna. A seguir, adotaram-se índices de precisão e revocação da indexação obtida, visando mensurar os resultados alcançados, para, adiante, registrar os procedimentos realizados em fluxos metodológicos para fins de replicação e consolidação da memória dos processos empregados. Este estudo diferencia-se dos demais pela sua capacidade de operar um tesauro de CI para padronização dos termos e abarcar técnicas de aprendizado de máquina para a tradução de linguagem natural em vocabulário controlado.

Com base nesta configuração, o objetivo geral desta tese é **propor um modelo metodológico para a formulação de indicadores temáticos de informação científica, a partir da integração teórica e metodológica entre a Indexação Automática e os Estudos Métricos da Informação**. Para tal, elaboraram-se os objetivos específicos mencionados a seguir:

a. Identificar os pontos de integração teórica e metodológica entre a IA e os EMI;
b. Delinear a trajetória metodológica de formulação de IT de informação científica;
c. Validar os resultados obtidos no percurso metodológico por meio das análises dos indicadores temáticos de informação científica presentes nos *corpora*.

A justificativa deste estudo para a área de CI e para o Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Pernambuco (PPGCI/UFPE), é colaborar com o desenvolvimento da linha de pesquisa “Comunicação e Visualização da Memória”, que contempla aspectos metodológicos e técnicos aplicados à produção, à gestão, à organização, à recuperação e ao uso da informação. Tal conexão se estabelece ao se conceber a proposição de um percurso metodológico utilizando *softwares* de EMI e de IA para a classificação do

conhecimento científico e produção de estudos metacientíficos que objetivam compreender o mapa temático da CI. No mais, reforça-se a demanda presente na CI, configurada pelos seus 27 programas de pós-graduação e centenas de pesquisadores, que necessitam de instrumentos facilitadores de geração dos IT, visando atender ao grande volume textual produzido. Contudo, a metodologia proposta pode servir a diversas outras áreas com interesse na extração de conhecimento de textos científicos.

Percurso metodológico pode ser entendido como a trajetória realizada por um pesquisador para atingir os objetivos de sua pesquisa. Conforme destacam Leite e Andrade (2020), são processos que devem ser acautelados no rigor científico, dada a complexidade dos fenômenos discutidos pela ciência. Deve ser construído de forma cuidadosa, a fim de prover análises autênticas, originais e confiáveis, que se somem ao conhecimento científico de um campo específico. Dado o valor de determinados percursos metodológicos para a construção de conhecimentos científicos críveis, optou-se por tornar o percurso metodológico o propósito desta análise, haja vista a possibilidade de beneficiar pesquisadores que desejam utilizar, criticar e replicar este modelo e seus respectivos processos. Nesta pesquisa, o percurso metodológico basear-se-á nas potencialidades de sinergia entre a IA e os EMI, e suas aplicações no contexto científico.

Outro fator de destaque, que justifica a elaboração desta pesquisa, são os ganhos para os diversos estudos, inclusive, métricos, com a construção de um procedimento apurado de análise temática que amplie as capacidades de produção científica dos pesquisadores de diversas áreas do conhecimento. Sendo assim, a CI pode contribuir de forma relevante, auxiliando no processo de descoberta de temas e assuntos em trabalhos científicos publicados em diferentes bases de dados. Tal característica contribui para a construção da memória científica da informação, a partir da identificação e visualização temática do status científico da área.

Outro ponto relevante é a necessidade de estimular estudos que promovam o acesso à informação de maneira a compreender a dinâmica da produção de conhecimento científico. Alguns trabalhos da CI já se utilizaram de metodologia semelhante e buscaram construir um panorama temático sobre determinado campo científico por meio das palavras-chave de forma manual. A exemplo disto, Ferreira (2012) elaborou um estudo propondo a análise manual das palavras-chave dos artigos publicados em oito Grupos de Trabalho (GTs) do Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação (Enancib), em 6 anos específicos.

Além desta pesquisa, pode-se citar o trabalho de Viana (2016), que objetivou identificar as temáticas das teses dos programas de pós-graduação em CI. O universo analisado foi de 52 teses de dois programas escolhidos. A autora utilizou o Tesauro Brasileiro de Ciência da Informação (TBCI) para classificar manualmente os documentos, em seguida, adotou a Taxonomia da Ciência da Informação de Donald T. Hawkins para realizar a classificação temática das teses. Ambos os estudos foram realizados de forma manual incorporando categorias previamente identificadas pelo indexador humano.

Portanto, o desenvolvimento de pesquisas que possibilitem ampliar a capacidade e a velocidade de categorizar tematicamente um conjunto de documentos científicos torna-se justificável e relevante. Compreende-se a necessidade em atribuir conceitos e facilitar o entendimento sobre determinados agrupamentos documentais com o propósito de facilitar a sua indexação e consequente recuperação da informação. Deste modo, esta pesquisa busca trilhar esse caminho com o propósito de se tornar um instrumento contributivo para a ciência brasileira.

Dentre as aplicações possíveis, cita-se o contexto de acelerada produção de conhecimento científico sobre o novo coronavírus, causador da Covid-19, motivado pela busca incansável por pesquisas que possam apontar soluções profiláticas e de tratamento contra a doença. Esses estudos estão sendo realizados em diversos laboratórios e universidades ao redor do mundo e têm resultado numa vasta quantidade de artigos científicos produzidos. Numa rápida consulta realizada no dia 21/10/2021, às 22:30h (UTC -3:00), na base de dados científicas da área de saúde, PubMed, com o termo “Covid-19”, foram encontrados 189.197 resultados, no qual, cerca de 189.015, foram publicados apenas desde 2020. Portanto, imagina-se que, em meio à leitura dos quase 189 mil artigos recuperados, se levaria muito tempo para encontrar informações cruciais que precisassem ser verificadas e analisadas de maneira rápida, e essa é a questão que o estudo aqui proposto pretende responder, promovendo a busca por formas alternativas de encontrar o conhecimento num amplo conjunto de textos científicos, simbolizando enorme avanço na pesquisa.

Outro fator relevante é a utilização de ferramentas que podem ser adotadas para facilitar o processo de análise da informação. Na IA, os recursos de automatização colaboram para a recuperação/extração de termos, enquanto as dos EMI buscam produzir indicadores de qualidade sobre os registros. Não obstante, é necessário que essas técnicas sejam avaliadas de acordo com sua capacidade de processamento e de resultados atingidos, principalmente para textos científicos em português do Brasil. Neste caso, compele-se a urgência do

desenvolvimento de estudos que possam propor metodologias adaptadas à realidade científica brasileira.

Ademais, enxergam-se as possibilidades de progresso do campo da Organização da Informação e do Conhecimento no Brasil, que tem nos processos de indexação importante força motriz. Sobre isto, os processos que envolvem a indexação podem proporcionar a construção de uma lista de representações dos documentos publicados. Por sua vez, esses documentos podem dispender tempo e dedicação daqueles que atuam no processo de Indexação Manual, pois dependendo do volume de textos a serem lidos e compreendidos, o processo pode se tornar desafiador. Por esse motivo, justifica-se a IA como aprimoramento da IM, tornando o processo mais rápido e menos custoso para todas as instituições e indivíduos envolvidos, atendendo, especialmente, o trabalho de revistas científicas, produtores e publicadores de pesquisas.

Isto posto, parte-se dos pressupostos explicitados a seguir: a) o processo de IA associado aos EMI auxilia na validação da coleta de vários IT, quando se trabalha com grandes volumes de textos, permitindo a identificação de descritores e conceitos; b) a dispersão terminológica das palavras-chave informadas nos registros bibliográficos é um fator limitante da realização de estudos métricos sobre as temáticas; c) os sistemas de IA podem ser utilizados como instrumentos na proposição de um percurso metodológico para a representação de temáticas; d) a IA permite maior rapidez no que tange à atribuição de termos em documentos, com isso, ela torna-se vantajosa, pois os custos envolvidos em contratação e treinamento de pessoas para realizá-la manualmente, muitas vezes, pode se tornar inviável, justificando, assim, a adoção de uma técnica para viabilizar os processos de indexação. Conforme afirma Medelyan:

> a importância da indexação automática de assuntos é evidente. Nas bibliotecas e em qualquer repositório centralizado de documentos, a indexação automática levaria uma carga significativa dos ombros dos bibliotecários. Na web, as ferramentas de sugestão de tags orientariam os usuários a documentos mais úteis. No processamento de linguagem natural, as frases-chave atribuídas automaticamente forneceriam uma dimensão semântica altamente informativa na representação de documentos que beneficiaria novas aplicações (MEDELYAN, 2009, p. 3, tradução nossa).[1]

[1] (MEDELYAN, 2009, p.3-4) The importance of automatic topic indexing is evident. In libraries and any centralized document repositories, automatic indexing would lift a significant burden from librarians' shoulders. On the web, tag suggestion tools would guide users to more useful documents. In natural language processing, automatically assigned keyphrases would provide a highly informative semantic dimension in document representation that would benefit new applications.

Em relação à IA, observa-se que ela evitaria a subjetividade inerente ao elemento humano na realização do processo. Gil-Leiva (1997, p. 2) atestou que o grau de coincidência entre os termos utilizados por indexadores humanos profissionais oscila entre 30% e 60%, indicando um alto percentual de inexatidão nos termos escolhidos, como foi afirmado:

> A subjetividade está presente no processo de indexação. O grau de concordância entre os termos de indexação atribuídos por diferentes indexadores profissionais é normalmente entre 30% e 60%. Sobre esses e outros aspectos do manifesto de Cleverson (1984), ao afirmar que se dois indexadores especialistas analisam o mesmo documento separadamente, eles convergem apenas em 30% dos termos propostos; se duas pessoas ou grupos constroem um tesauro, eles concordam apenas com 60% dos termos incluídos; se dois profissionais interrogam um banco de dados com a mesma pergunta, apenas 40% das informações recuperadas são comuns; e, por fim, se dois cientistas forem questionados sobre a relevância de um conjunto de documentos, para uma determinada questão, a concordância entre eles não ultrapassa 60% (Fator de subjetividade) (GIL-LEIVA, 1997, p. 2, tradução nossa).[2]

Portanto, observando os aspectos levantados pelo autor, a subjetividade humana pode ser um fator limitante na capacidade de escolha de documentos e termos analisados num determinado conjunto de textual ou base de dados.

Nesse sentido, acredita-se, a partir do ponto de vista desta pesquisa, que a IA pode ser útil no desenvolvimento de IT, pois desse modo a grande quantidade de dados gerados nos mais diversos ambientes informacionais pode ser tratada por *softwares* com alto grau de confiabilidade. Segundo Simões *et al*. (2017), os dados de produção digital alcançaram a quantidade de 44 bilhões de terabytes, em 2020, o que corresponde a uma quantidade gigantesca de dados para que um ser humano possa acessar em toda sua vida. Os mesmos autores ainda afirmam que, apesar da evolução constante dos Sistemas de Recuperação da Informação (SRI), existe uma dificuldade no que diz respeito às buscas baseadas em temas ou conceitos e, consequentemente, os índices de revocação e precisão ficam abaixo do esperado. Sendo assim, a presente pesquisa justifica-se por buscar a facilitação da representação dos registros bibliográficos, uma vez que utiliza a rapidez do processamento de máquinas, concedendo maior acesso aos conteúdos produzidos, em menor tempo.

[2] La subjetividad está presente en el proceso de la indización. El grado de coincidencia entre los términos de indización asignados por indizadores profesionales diferentes suele oscilar entre el 30% y 60%. Sobre estos y otros aspectos de manifestó Cleverson (1984) cuando expresó que sí dos indizadores expertos analizan separadamente un mismo documento sólo convergen en el 30% de los términos propuestos; si dos personas o grupos construyen un tesauro solamente concuerdan en el 60% de los términos incluidos; si dos profesionales interrogan una base de datos con la misma cuestión sólo el 40% de la información recuperada es común; y por último, si se pregunta a dos científicos sobre la relevancia de un conjunto de documentos, para una determinada cuestión, el acuerdo entre ambos no excede del 60%. (Factor subjetividad). (GIL-LEIVA, 1997, p. 2)

Isto posto, vale ressaltar que o produto mais importante dessa pesquisa é a proposição de um percurso metodológico que possibilite a visualização de um conjunto de indicadores temáticos a serem analisados de maneira consistente e com baixa incidência de erros.

Para tanto, assume-se o rigor do processo avaliativo da ciência, com os critérios de qualidade e quantidade, como afirmaram Vanz e Stumpf (2010, p.67):

> Quanto mais ativo e produtivo o ambiente científico, mais freqüentes e rigorosas são as rotinas de avaliação vigentes. Estes processos avaliativos se fundamentam, principalmente, em duas metodologias: a avaliação qualitativa, feita pelos pares, fortemente ancorada na reputação adquirida pelo avaliado; e a que se deriva de critérios quantitativos, baseados em métodos bibliométricos e cientométricos.

Visando reconhecer as ferramentas utilizadas no cenário internacional e valorizar os pesquisadores que contribuem para a evolução do campo da IA, é interessante analisar a trajetória da produção científica neste campo. A CI, enquanto ciência neófita, é capaz de projetar e utilizar ferramentas de bases de dados nacionais e internacionais, fornecendo uma quantidade de informações adequada ao fornecimento da criatividade à pesquisa. Os méritos da busca e da Recuperação da Informação (RI) são aspectos importantes da CI e se destacam por trabalhar com documentos presentes nas bases científicas, contribuindo para a evolução da ciência brasileira. Segundo Bufrem, Silva e Sobral (2017, p. 116):

> enquanto área do conhecimento científico, a Ciência da Informação (CI) no Brasil é composta por elementos que a caracterizam como espaço de produção de saberes consolidado, buscando atender as demandas de cunho acadêmico e profissional que circundam o universo da informação através de práticas disciplinares e interdisciplinares. Dentre os elementos constituintes da CI, destacam-se os formativos, composto por programas de graduação e pós-graduação que promovem o ensino, a pesquisa e a extensão na área; os representativos, que englobam as associações, os modos de comunicação e os eventos que discutem a organização comunitária e a representatividade do campo em esfera social; e por fim, os elementos avaliativos, que convergem para o fortalecimento das estruturas formais, institucionalização social do campo, regramento e adequação da área aos critérios de qualidade científica.

Desse modo, busca-se, por meio desta pesquisa, sondar o aspecto avaliativo e procurar respostas aos questionamentos propostos. Compreende-se, nesta investigação, a percepção de uma ciência que deseja investigar a qualidade das produções desenvolvidas e mensurar o nível das pesquisas a partir de dois enfoques: EMI e IA. Para tanto, a adoção de metodologias, como a dos IT e dos sistemas de IA, num processo contínuo, auxiliará às demandas dos campos científicos. Isto posto, compreende-se que a ideia de construir uma pesquisa nessa determinada ótica nasceu das dificuldades em analisar os campos científicos devido à ausência de controle terminológico na especificação dos metadados de assuntos armazenados

nas bases de publicações científicas. Tal descontrole dificulta, consequentemente, a busca e a RI, bem como os estudos métricos com IT.

Com o propósito de facilitar a compreensão sobre o trajeto e a construção desta tese, segue a descrição do que se pretende alcançar em cada seção. No primeiro capítulo, é apresentada a introdução; nela, expõe-se: o contexto em que se apresenta a pesquisa; a problemática e a pergunta norteadora; o objetivo geral e os objetivos específicos; e as justificativas, isto é, as motivações que impulsionaram o desenvolvimento desta pesquisa.

Na segunda parte, são elucidados os apontamentos teóricos e as relações conceituais sobre os temas aqui propostos. Nela, estão apresentadas a dimensão teórica e suas relações implícitas e explícitas com os objetos de estudo. Na terceira seção, há uma dissertação a respeito das etapas e dos procedimentos envolvidos na construção do trabalho. Nela, são evidenciados os aspectos classificatórios da pesquisa e se revelam, de forma sistemática, todas as etapas realizadas.

Na quarta seção, são explicitados os resultados da pesquisa e suas respectivas discussões. Na quinta, são expressas as considerações finais, contendo as principais impressões, dificuldades encontradas e sugestões de estudos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, apresentam-se as questões conceituais necessárias à fundamentação desta tese. Na seção 2.1, são discutidos os contextos e as definições dos EMI, bem como sua relevância para a proposta de percurso metodológico deste estudo. Em seguida, na subseção 2.1.1, tem-se uma discussão sobre a importância da ciência que é produzida nos diversos meios de comunicação científica, expondo pontos relevantes sobre questões inerentes à divulgação do conhecimento científico, e, na subseção 2.1.2, debatem-se os principais trabalhos identificados sobre Representação Temática por meio dos EMI. Ao final, na subseção 2.1.3, há um breve relato com os principais conceitos sobre o software de EMI utilizado nesta pesquisa.

Em seguida, isto é, na seção 2.2, reflete-se sobre conceitos relativos à Descoberta de Conhecimento (DC), especificamente, os relacionados à busca de determinadas temáticas em textos científicos. Posteriormente, na subseção 2.2.1, relata-se alguns exemplos de trabalhos que lidam com a Mineração de Textos (MT) que, por sua vez, tem como uma de suas tarefas a indexação. *A posteriori*, na seção 2.3, são apresentados os conceitos de IA e suas principais características. Na subseção 2.3.1, é realizado um estudo bibliométrico sobre IA trazendo algumas especificidades da área e pontuando autores mais produtivos e documentos mais representativos, em termos de quantidade de citações. Logo após, na subseção 2.3.2, são descritos alguns sistemas de IA utilizados no mundo, considerando os resultados obtidos na seção anterior. E por fim, na subseção 2.3.3, se discute sobre os principais conceitos relacionados ao *software* de IA adotado nesta pesquisa.

2.1 Estudos Métricos da Informação

A busca pelo conhecimento científico visa identificar, nos fatos supostamente inexplicáveis que ocorrem no universo, padrões e regras que regem suas ocorrências. Essas características podem ser expressas por meio de elementos descritivos, a exemplo da linguagem. Os símbolos mais frequentemente utilizados, de forma uníssona, pelas áreas científicas, são aqueles que permitem a contagem da quantidade de algum elemento a partir de um conceito pré-definido. A partir de uma dimensão mensurável, o indivíduo procura reconhecer sua realidade por meio da quantificação.

> O Universo é dinâmico, está permanentemente em mutação, e essa se dá de modo muito complexo e intrincado. Mas a mente humana é capaz de isolar

> algumas partes dessa mutação, particularmente aquelas que se repetem frequentemente e cujo desfecho é de importância para a sobrevivência ou o bem-estar – esses são os fenômenos ou processos. Entendê-los, com o propósito de conduzir seu desfecho sempre no sentido de um resultado almejado a priori (ou seja, de gerar uma tecnologia) é a origem do conhecimento científico, foi sua primeira e maior preocupação, e segue sendo uma das mais importantes na atualidade (TRZESNIAK, 2014, p. 6).

Considerando o que afirma Trzesniak (2014), pode-se conceber que o propósito da ciência é conhecer sua realidade de pesquisa e as diversas particularidades que estão atreladas a ela, identificando um padrão e buscando descrevê-lo de acordo com um método pré-existente que satisfaça todas as variáveis envolvidas na análise. Do contrário, o cientista não teria capacidade técnica para buscar quantificar uma realidade de pesquisa sem antes decifrar as origens das áreas do conhecimento que contribuem para estabelecer regras e padrões na forma como as informações são disponibilizadas e obtidas.

Nesse sentido, uma das disciplinas nucleares da CI é a Bibliometria, pois surge com o propósito de mensurar as diversas características presentes num determinado documento ou livro, desde a quantidade de páginas até à quantificação das citações e referências. As discussões sobre suas possíveis definições iniciam-se em torno do conceito atribuído por Pritchard, que afirma:

> Já em 1922, Hulme usava o termo "bibliografia estatística" em suas preleções na Universidade de Cambridge. Segundo o autor, ele empregava esse termo para designar o processo de esclarecer a história da ciência e da tecnologia pela consideração de documentos. Em anos ulteriores, Pritchard usou a palavra "bibliometria" a fim de descrever a análise quantitativa das citações. Enquanto que os historiadores da ciência russa sugeriram o uso do termo "cientometria" para semelhante tipo de estudo (PRITCHARD, 1969 apud GARFIELD, 1973, p. 137).

Percebe-se então que o termo "Bibliometria" buscou, inicialmente, quantificar e mensurar as citações de um texto; porém, hoje, suas características estão muito mais voltadas para outras medidas dos documentos científicos. A popularização do conceito se deu em meados de 1969, a partir da publicação de um artigo que possuía a pergunta: "Bibliografia Estatística ou Bibliometria?", como problemática. Contudo, a ideia passou a ser melhor compreendida posteriormente, na década de 70, quando começou a ser amplamente utilizada por aqueles que buscavam uma avaliação mais quantitativa da ciência (ARAUJO, 2006).

Um dos aspectos mais interessantes nos estudos bibliométricos, são as pesquisas que envolvem análise de citações. A citação é processo de mencionar ou citar determinado trabalho ou autor num texto, seja ele científico ou não, sendo assim, quantificar o número de citações em determinado documento, é campo de estudo alvo da bibliometria. No que tange

os estudos de citação, pode-se mencionar um trecho de Araújo (2018, p.45), em que o pesquisador afirma que o surgimento do primeiro índice de citação, o SCI (*Science Citation Index*), impactou positivamente o crescimento da subárea dos EMI. Ele também utiliza uma citação de Garfield (1978) para discorrer sobre o início de tais estudos:

> [...] Tal índice foi fruto de uma ideia surgida em 1955, e foram necessários oito anos para se formular as bases teóricas e conceituais da análise de citações, buscando fundamentações nos trabalhos de sociologia da ciência de Merton e de comunicação científica de Crawford, Griffith e Crane (GARFIELD, 1978 apud ARAÚJO, 2018, p. 45).

Outrossim, visualizam-se os Estudos de Citação como percussores de diversas teorias que buscam justificar as razões pelas quais um pesquisador cita um determinado autor em seu trabalho. Esses estudos fazem parte das discussões desenvolvidas no âmbito da comunicação científica, que estão brevemente comentadas no referencial teórico desta tese.

Tais estudos de citação baseiam-se em teorias que procuram apresentar a variedade dos modelos textuais existentes e das dinâmicas relacionadas aos contextos que motivam a ocorrência das citações, como afirmaram Silveira e Caregnato (2018, p. 60):

> É importante registrar, todavia, que o princípio norteador para a caracterização dos contextos que permeiam a atividade científica se operacionaliza pelas conexões que os atores realizam entre os significados das práticas sociais e as situações que determinam as práticas de citação.

Desta feita, as razões que levam determinado autor a realizar uma citação ou apontar determinado contexto podem ser averiguados por disciplinas dos EMI como a bibliometria, por exemplo. Um autor que discute bem essa relação é Okubo (1997, p. 9), que também discute os conceitos de Bibliometria. Ele afirma ser esta uma área de estudo multidisciplinar, aplicável a uma grande variedade de domínios, que estuda a historicidade da ciência e o desenvolvimento de disciplinas científicas por meio do acompanhamento de movimentos históricos, os quais revelam-se através dos resultados alcançados pela comunidade científica.

Examinando a literatura científica, pode-se afirmar que a Bibliometria se relaciona à Sociologia da Ciência, realizando análises da comunidade científica e a estrutura de uma dada sociedade, revelando aspectos às redes de pesquisadores e suas respectivas motivações. Ela ainda estuda a documentação mediante à contagem do número de periódicos existentes, podendo identificar o núcleo e a periferia da produção acadêmica. E, ainda, possui poder sobre a política científica, interessando-se por indicadores de produtividade e de qualidade científica e tecnológica (OKUBO, 1997).

Da mesma forma, Fonseca (2015) procura atribuir ao desenvolvimento da pesquisa um aspecto de mensuração, fazendo com que se torne algo presente na quantificação, pela

necessidade de construir uma certeza científica, renovando a epistemologia das ciências humanas e afastando a ideia de que apenas o método unicamente qualitativo é essencial. Para o autor, a estatística é conclusiva, e seria, a partir de dado momento, muito mais generalizante, a depender dos modelos construídos.

Já Araújo (2018, p. 73) concebe que:

> Os estudos métricos historicamente privilegiam a produção científica e sempre desenvolveram pesquisas buscando medir índices principalmente a partir de citações com o objetivo de avaliação de instituições, de produtividade de autores e para ranqueamento de revistas, entre outros.

Urbizagástegui Alvarado (1984, p. 91) afirma que a Bibliometria se originou de outras áreas, como a Sociologia e a Psicologia; considerando, também, que segmentos de áreas do conhecimento como a Psicometria, utilizada pela Psicologia, a Econometria, componente da Economia, e a Sociometria surgem no início do século XX, num processo de medir o conhecimento realizando estudos matemáticos e utilizando técnicas experimentais baseadas em métodos quantitativos. Por conseguinte, o autor concebe que a Sociometria acaba sendo replicada em outras áreas, a exemplo da Educação e da Administração, até que se chegue na Biblioteconomia.

Consequentemente, com a adoção dessas áreas pela Biblioteconomia, teorias e leis foram criadas para compor a análise dos fenômenos específicos deste campo. O quadro abaixo apresenta as principais leis e estudos que foram adotados pela área.

Quadro 1- Principais Leis e Estudos da Bibliometria

LEIS E ESTUDOS	OBJETIVO PROPOSTO
Lei de Bradford (1934)	Descrever a distribuição da literatura periódica numa área específica.
Lei de Lotka (1926)	Apresentar o índice de produtividade dos autores.
Lei de Zipf (1949)	Detalhar a frequência no uso de palavras.
Lei de Goffmam e Newill (1967)	Descrever a difusão da comunicação escrita como um processo epidêmico.
Lei do Elitismo de Price	Relatar como uma pequena parte da literatura mais recente está relacionada a uma parte maior da literatura mais antiga.
Lei de Obsolescência	Expor a queda da validade ou utilidade de informações no decorrer do tempo.

Fonte: Adaptado de (URBIZAGÁSTEGUI ALVARADO, 1984, p. 91).

Dentre as principais Leis apresentadas, destacam-se as Leis de Bradford, Lotka e Zipf, que são as mais utilizadas na CI e em diversos estudos bibliométricos desenvolvidos.

Urbizagástegui Alvarado (2002, p. 14) define a Lei de Lotka como "o número de autores que fazem **N** contribuições em um determinado campo científico é aproximadamente $1/n^2$ daqueles que fazem uma só contribuição, e que a proporção daqueles que fazem uma única contribuição é de mais ou menos 60%". Essa Lei busca comprovar que a quantidade de contribuições, isto é, artigos publicados, geralmente em 60% dos casos, é de uma contribuição por pesquisador para determinada área, ou seja, 60% contribuem apenas uma vez.

A segunda Lei Bibliométrica, conhecida como Lei de Bradford, pode ser descrita assim:

> Se dispormos periódicos em ordem decrescente de produtividade de artigos sobre um determinado tema, pode-se distinguir um núcleo de periódicos mais particularmente devotados ao tema e vários grupos ou zonas que incluem o mesmo número de artigos que o núcleo, sempre que o número de periódicos existentes no núcleo e nas zonas sucessivas seja de ordem de 1: n: n2 : n3 .... Assim, os periódicos devem ser listados com o número de artigos de cada um, em ordem decrescente, com soma parcial. O total de artigos deve ser somado e dividido por três; o grupo que tiver mais artigos, até o total de 1/3 dos artigos, é o "core" daquele assunto. O segundo e o terceiro grupo são as extensões. A razão do número de periódicos em qualquer zona pelo número de periódicos na zona precedente é chamada "multiplicador de Bradford" (Bm): à medida que o número de zonas for aumentando, o Bm diminuirá (ARAÚJO, 2006, p. 15).

Esta Lei Bibliométrica foi muito importante à época, pois possibilitava identificar, dentre os diversos periódicos existentes, aqueles que apresentavam a maior incidência de discussão, em seus artigos indexados, sobre um determinado tema. Dessa maneira, ao se atribuir uma política de aquisição numa unidade de informação, era mais simples decidir qual periódico adquirir, a depender das necessidades do usuário e da instituição.

Na mesma proporção, a Terceira Lei Bibliométrica, a Lei de Zipf, é descrita por Araújo (2006, p. 16) como:

> A terceira das leis bibliométricas clássicas é a Lei de Zipf, formulada em 1949 e que descreve a relação entre palavras num determinado texto suficientemente grande e a ordem de série destas palavras (contagem de palavras em largas amostragens). Zipf, analisando a obra Ulisses de James Joyce, encontrou uma correlação entre o número de palavras diferentes e a freqüência de seu uso e concluiu que existe uma regularidade fundamental na seleção e uso das palavras e que um pequeno número de palavras é usado muito mais freqüentemente. Ele descobriu que a palavra mais utilizada aparecia 2653 vezes, a centésima palavra mais utilizada ocorria 256 vezes e a duocentésima palavra ocorria 133 vezes. Zipf viu então que a posição de uma palavra multiplicada pela sua freqüência era igual a uma constante de aproximadamente 26500. Sua proposta, assim, é de que, se listarmos as palavras que ocorrem num texto em ordem decrescente de freqüência, a posição de uma palavra na lista multiplicada por sua freqüência é igual a uma constante. A equação para esse relacionamento é: r x f = k, onde r é a posição da palavra, f é a sua frequência e k é a constante.

Apesar de ser duramente criticada pelos estudos recentes, a relação descoberta por Zipf aponta para a existência de uma economia no uso de palavras pelos pesquisadores e para o fato de que, na maioria das vezes, as palavras mais frequentes indicam o assunto do documento. Essa Lei é interessante pois fornece uma das explicações estatísticas para que os *softwares* de IA, por exemplo, escolham termos mais frequentes, após a remoção das *stopwords* como sendo o que representa o conteúdo semântico do documento. Especificamente esta lei, representa o ponto de partida das justificativas teóricas adotadas na escolha do percurso metodológico representado nesta investigação.

Outro ponto relevante, aponta que grande parte das teorias desenvolvidas para examinar o comportamento da ciência possui como cerne a mensuração de algum objeto ou a quantificação de um fenômeno. Tal aspecto é relevante pois justifica a evolução da estatística e da probabilidade como campos que podem contribuir para as análises dos comportamentos da ciência.

Entre os especialistas na área dos EMI pode-se citar um trabalho de Macias-Chapula (1998). Em seu estudo, o pesquisador utiliza algumas definições para Bibliometria, Cienciometria e Informetria do estudioso Jean Tague-Sutcliffe. Conforme Tague-Sutcliffe (1992, apud Macias-Chapula, 1998, p. 134), as 3 disciplinas supracitadas podem ser classificadas de acordo o quadro 2.

Quadro 2 – Definições das 3 principais disciplinas inseridas nos EMI

DISCIPLINAS	DEFINIÇÕES
BIBLIOMETRIA	Estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada. A Bibliometria desenvolve padrões e modelos matemáticos para medir esses processos.
CIENCIOMETRIA	Estudo dos aspectos quantitativos da ciência enquanto uma disciplina ou atividade econômica. Envolve estudos quantitativos das atividades científicas incluindo publicações.
INFORMETRIA	Estudo dos aspectos quantitativos da informação em qualquer formato. Pode incorporar, utilizar e ampliar os muitos estudos de avaliação da informação que estão fora dos limites da Bibliometria e da Cienciometria.

Fonte: (MACIAS-CHAPULA, 1998, p. 134).

Os conceitos citados acima reforçam que o domínio dos EMI possui diversas possibilidades de abordagem dependendo muito mais dos objetos a serem analisados. Entretanto, vem surgindo e evoluindo outras correntes dos estudos métricos. Essas correntes classificam os estudos em outros níveis e categorias que evoluem de acordo com as diferenças dos objetos de estudo analisados. O quadro 3 reforça os conceitos mencionados no quadro 2, além de ampliar a descrição das técnicas existentes:

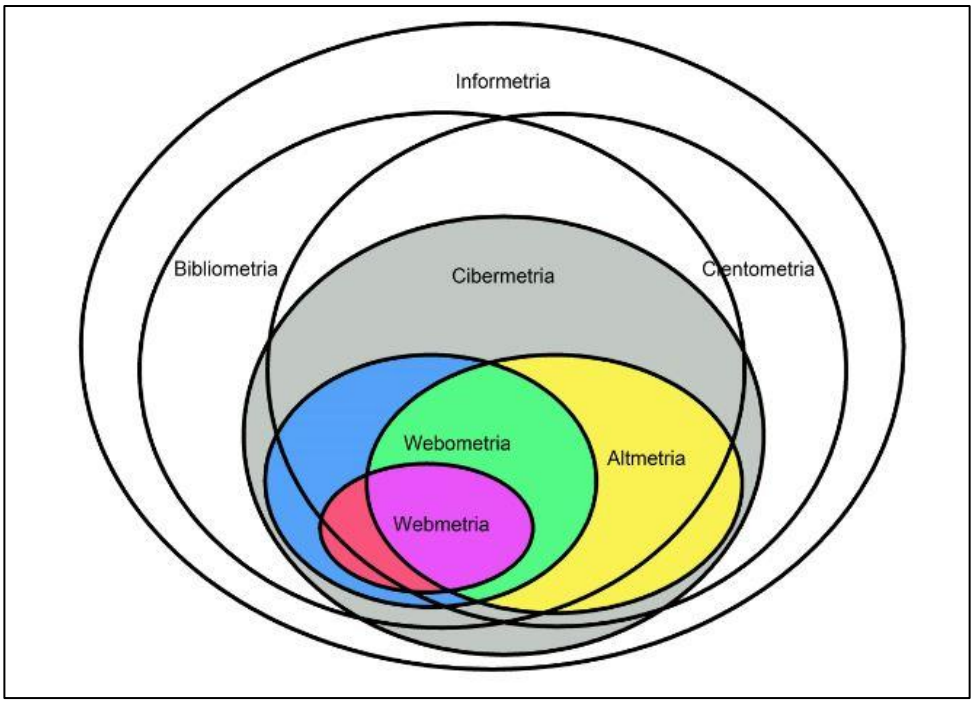
Quadro 3 – Finalidades e Objetos das disciplinas inseridas nos EMI

TÉCNICAS	FINALIDADE	OBJETOS DE ESTUDO
BIBLIOMETRIA	Estudo dos aspectos quantitativos da produção, disseminação e uso da informação registrada. A Bibliometria desenvolve padrões e modelos matemáticos para medir o uso de documentos e organização de serviços bibliográficos.	Documentos (livros, artigos, teses, dissertações, autores, usuários)
CIENCIOMETRIA	Estudo dos aspectos quantitativos da ciência enquanto uma disciplina ou atividade econômica. Envolve estudos quantitativos das atividades científicas identificando domínios de interesse.	Disciplinas, campos, áreas, assuntos específicos.
INFORMETRIA	Estudo dos aspectos quantitativos da informação em qualquer formato. Envolve os estudos de medição dos sistemas de informação e de recuperação da informação.	Palavras, documentos, bases de dados.
BIBLIOTECOMETRIA	Estudos dos aspectos quantitativos da gestão, serviços e organização de bibliotecas e acervos.	Bibliotecas
WEBMETRIA	Estudos dos aspectos quantitativos da organização e uso de sites.	Páginas da internet, hospedeiros de sites.
PATENTOMETRIA	Estudos dos aspectos quantitativos da produção de atividade tecnológica por meio de inovações.	Patentes
ALTMETRIA	Estudos dos aspectos quantitativos dos índices de acesso, comentários, citações e menções de pesquisas publicadas em redes sociais.	Comentários, citações, compartilhamento de links de trabalhos científicos.

Fonte: Adaptado de (NORONHA; MARICATO, 2008) e (GOUVEIA, 2013)

De acordo com Noronha e Maricato (2008), os autores analisam a evolução dos EMI de acordo com o campo de estudo, a metodologia e a tecnologia utilizadas, além das abordagens e variáveis que são analisadas em cada técnica. Percebe-se que a fundamentação para a construção das categorias propostas evolui das ideias apresentadas por Macias-Chapula (1998) e Sanz Casado (2006). Os autores atribuem às disciplinas a denominação de técnicas e justificam essa classificação pela diferença entre suas finalidades e objetos de estudo. Adicionalmente ao quadro 3, pode-se identificar a mais recente das técnicas que estão sendo utilizadas pelos o EMI, a Altmetria. Essa técnica é mencionada amplamente pelo autor Gouveia (2013) no qual ele elabora uma distinção entre as outras técnicas e reconhece um novo campo de estudo para os EMI, em sua pesquisa o autor representa visualmente essa divisão na figura 1:

Figura 1 – Divisões entre as disciplinas no campo dos EMI



Fonte: GOUVEIA (2013, p. 221)

A figura 1 consegue representar como as técnicas se intercalam e se relacionam, observa-se que as mais recentes surgem como subdivisões das grandes áreas da Informetria, Bibliometria e da Cientometria. O Autor cita ainda outras que não foram mencionadas por Noronha e Maricato, como a Cibermetria e a Webometria. Cibermetria surge como uma junção dos termos Cibernético e Métricas, ou seja, os elementos que intercalam a internet, mundo digital e os estudos métricos. Assim como a Webometria, num aspecto mais específico, une os elementos da Web e dos EMI. Assim sendo, é possível compreender a

complexidade das áreas e justifica o comportamento subordinado às áreas maiores que promoveram as suas origens.

Em contrapartida, verifica-se que a Cientometria e a Bibliometria concentram-se em áreas bem definidas, as primeiras a surgirem como subdivisões da Informetria, elas tornaram-se disciplinas fundamentais e de grande importância metodológica, e concentram diversas outras disciplinas, como pode ser observado a seguir:

Quadro 4 – Áreas de concentração da Bibliometria e Cienciometria

1. Aspectos estatísticos da linguagem e frequência de citação de frases, tanto em textos (linguagem natural), como em índices impressos e em formato eletrônico.
2. Características da relação autor produtividade medidas por meio do número de artigos ou outros meios, além de mensurar o grau de colaboração.
3. Características das publicações, sobretudo a distribuição em revistas de artigos relativos a uma disciplina.
4. Análise de citação: distribuição entre autores, artigos, instituições, revistas, países; uso em avaliação; mapa de disciplinas baseado na co-citação.
5. Uso da informação registrada: circulação em bibliotecas e uso de livros e revistas da própria instituição; uso de bases de dados.
6. Obsolescência da literatura, avaliada pelo uso e pela citação.
7. Crescimento de literaturas especializadas, bases de dados, bibliotecas; crescimento simultâneo de novos conceitos.
8. Definição e medição da informação.
9. Tipos e características dos níveis de desempenho da recuperação.

Fonte: Adaptação de (MACIAS-CHAPULA, 1998, p. 135).

Todos os aspectos enumerados acima apontam para a necessidade de quantificar e mensurar o que é produzido e como tal processo ocorre. Além disso, é possível calcular tendências no surgimento e no crescimento de novos conceitos, a exemplo da observação de quais autores são mais representativos em determinados campos e como as áreas se relacionam para que seja produzida ciência. Analisando pela ótica das áreas de concentração identificadas, entende-se que a um e a oito se aproximam com maior clareza desta pesquisa, uma vez que se busca propor um percurso metodológico, percorrendo a trajetória da produção científica presente em registros bibliográficos por meio de temáticas.

De forma mais abrangente, busca-se por meio da informetria medir palavras em documentos científicos, dessa forma, a bibliometria e cientometria se entrelaçam como técnicas e disciplinas que promovem esse tipo de análise quando possuem como objeto o documento. Neste sentido, essa tese busca construir os elementos necessários para elucidar as temáticas abordadas num conjunto definido de documentos.

De acordo com Marcelo e Hayashi (2013, p. 143) a Bibliometria compreende:

> A utilização da análise bibliométrica em pesquisas científicas se pautam na investigação do comportamento do conhecimento e da literatura como parte dos processos de comunicação. Embora a Bibliometria tenha sua maior aplicação nos campos da Ciência da Informação, é possível aplicá-la em várias áreas do conhecimento a fim de explorar o impacto da produção de um determinado campo de conhecimento, a produção e produtividade de um conjunto de investigadores, por meio da construção de indicadores bibliométricos.

À luz das ponderações acima realizadas, faz-se compreensível que a forma como a análise bibliométrica é utilizada na CI exerce influência na capacidade de observar o comportamento da ciência, de acordo com os impactos da produção acadêmica; sendo esta um produto ativo da atuação de pesquisas financiadas pelos órgãos institucionais, que auxiliam no desenvolvimento de estudos por meio do pagamento de bolsas de fomento. Nesse contexto, a ciência, em muitos momentos, depende da vinculação às regras exigidas pelos órgãos superiores, que, por sua vez, acabam ditando a forma como ela deve se comportar diante dos meios de avaliação e monitoramento.

2.1.1 Análise temática e suas relações

A investigação científica pode ocorrer de acordo com as demandas existentes em uma determinada realidade. No Brasil, o *modus operandi* da análise da produção científica possui relação direta com os estudos desenvolvidos nas universidades e grupos de pesquisa que trabalham com esta temática. Parte significativa dos pesquisadores brasileiros que trabalham com EMI está vinculada à área da CI, nas Universidades Federais, como afirmaram Alvarez e Caregnato (2017, p.23):

> A Ciência da Informação, a partir da denominada “explosão da informação” e do desenvolvimento das novas tecnologias, contribuiu notoriamente para a ampliação das avaliações quantitativas da produção do conhecimento científico. A criação de bases de dados como, por exemplo, a Web of Science, foi fundamental para a realização de estudos e a obtenção de indicadores bibliométricos. A formulação de leis e teorias ajudou a compreender o comportamento e a composição da literatura publicada nos periódicos das diferentes áreas.

Como foi observado pelos autores, a avaliação da produção científica evoluiu muito com a CI e depende cada vez mais das pesquisas realizadas por pesquisadores dessa área. Com o intuito de levantar esses dados, eles dependem de um esforço contínuo na aquisição de ferramentas e softwares que possam avaliar a ciência. Dessa maneira, os estudos ficam submetidos às limitações inerentes às ferramentas utilizadas, de acordo com os projetos que

são aprovados pelas instituições, e pelos incentivos financeiros dos órgãos governamentais como a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Consequentemente, o pesquisador acaba escolhendo como objeto de pesquisa os instrumentos que são possíveis de ser adquiridos, influenciando a amplitude e o alcance do estudo a que ele se propôs desenvolver.

Além disso, a identificação dos temas utilizados pelos pesquisadores em seus trabalhos não é trivial, apesar da existência de grandes bases de dados de pesquisadores como a plataforma Lattes[3], é sem dúvidas um grande desafio investigar os tópicos trabalhados, identificando os aspectos complexos resultantes do andamento de diversas pesquisas. Algumas iniciativas, como a criação de sistemas de extração de metadados de currículos, como o Scriptlattes, e-lattes, por exemplo, que buscam extrair as informações das publicações e conteúdos publicados na plataforma são promissoras, porém, não é simples a tarefa de mapear e construir os indicadores sobre as temáticas necessárias.

Outro problema considerável e que dificulta uma análise mais profunda do que é produzido nos documentos científicos, é a dificuldade na padronização dos termos encontrados nos documentos, utilizados como descritores ou palavras-chave, como relata Santos (2015, p. 641):

> A dispersão decorrente da variabilidade e granularidade de assuntos abordados em um conjunto de artigos de periódicos é grande quando analisamos as palavras-chaves de cada artigo, além de ocorrerem inconsistências, uso de sinônimos, variações linguísticas, bem como critérios diferenciados de nivelamento do tratamento da informação pelos autores, principalmente em áreas interdisciplinares, o que torna importante a discussão sobre a bibliometria temática e gera a necessidade de analisar os procedimentos adotados.

Outro ponto relevante, no que diz respeito à ausência de controle terminológico, é a ausência de padronização nos títulos dos periódicos científicos onde são indexados os artigos e nos nomes dos autores dos trabalhos. Essa dificuldade pode resultar na inconsistência dos dados coletados tornando as análises dos indicadores mais complexas, como relatam Pinto e Matias (2011, p. 13):

> Erros ou ausência de padrões na descrição de títulos de periódicos ou de autores, por exemplo, podem gerar informações imprecisas que tornam os indicadores distorcidos e não confiáveis para fundamentar análises e processos de tomada de decisão. Estas variáveis demandam precisão, pois

[3] Base de dados de currículos de pesquisadores, grupos de pesquisa e instituições ligadas à Ciência, Tecnologia e Inovação. (http://lattes.cnpq.br/)

> compõem o registro do conhecimento científico e, portanto, envolvem canais que fazem a conexão entre o que é efetivamente produzido nas universidades brasileiras e o que pode ser detectado e percebido como produção científica pelos gestores e pela sociedade em geral.

Desta forma, torna-se relevante a discussão dos indicadores precisos, quando eles se propõem a analisar qualitativamente a produção científica, utilizando abordagens que buscam realizar a representação de um domínio a partir da elaboração de mapas temáticos.

Outrossim, a existência de indicadores responsáveis por identificar os temas desenvolvidos nas pesquisas é uma tarefa fundamental no processo de descoberta do conhecimento nos documentos publicados pelos pesquisadores. De acordo Mugnani, Jannuzzi e Quoniam (2004, p. 124) "[...] os indicadores de CT&I são medidas quantitativas que buscam representar conceitos muitas vezes intangíveis dentro do universo do fazer ciência e da tecnologia [...]". Dessa forma, um indicador com essa característica intangível, é um desafio, quando se propõe a estabelecer uma medida quantitativa, usada para dimensionar elementos que serviriam de avaliação e identificação do progresso científico e tecnológico.

Kobashi e Santos (2006) relatam a existência de um conjunto expressivo de indicadores empregados na análise da produção científica, que podem ser divididos em indicadores de produção científica, de citação e de ligação, conforme o Quadro 5.

Quadro 5 - Indicadores e utilidades para a análise da produção científica

INDICADORES	UTILIDADES
DE PRODUÇÃO	Construídos pela contagem do número de publicações por tipo de documento (livros, artigos, publicações científicas, relatórios etc.), por instituição, área de conhecimento, país, dentre outros;
DE CITAÇÃO	Estabelecidos pela contagem do número de citações recebidas por uma publicação de artigo de periódico. É o meio mais reconhecido de atribuir crédito ao autor;
DE LIGAÇÃO	Criados pelas **co-ocorrências** de autoria, citações e **palavras**, sendo aplicados na **elaboração de mapas de estruturas de conhecimento** e de redes de relacionamento entre pesquisadores, instituições e países. Emprega técnicas de análise **estatística de agrupamentos**.

Fonte: Adaptado de Kobashi e Santos (2006)

Como pode ser observado no quadro 5, as utilidades dos indicadores variam de acordo com sua função, mais especificamente os de ligação, que tratam das coocorrência de palavras, são o principal alvo desta pesquisa, com a elaboração de indicadores temáticos dos resultados

obtidos da indexação automática dos documentos. Como foi afirmado pelos autores, esses indicadores podem apresentar padrões de comportamento nos campos científicos, identificando como o conhecimento se disseminou ao longo do tempo e suas possíveis relações e dinâmicas.

Os autores afirmam ainda que a análise dos campos dos descritores em bases de dados é um caminho possível, pois o mapeamento temático dos vocabulários controlados pela própria fonte de dados seria mais seguro. Esse ponto de vista, leva em consideração que a indexação realizada nessas bases, utilizando-se de tesauros, forneceria informações temáticas mais padronizadas do que as palavras de um texto completo. Neste sentido, "os mapas gerados a partir desses dados e métodos são representações da produção científica da área expressa por meio de conceitos produzidos e utilizados pela própria área". Consequentemente, a segurança das terminologias utilizadas na área seria importante na geração das representações temáticas (KOBASHI; SANTOS, 2006, p. 34).

Outra questão relevante diz respeito ao formato que são elaboradas e construídas as análises temáticas, que muitas vezes se utilizam de classificação por assuntos gerais ou específicos, por meio de linguagem natural ou documentária. Neste caso, podem utilizar a indexação existente na base de dados consultada, exigindo, muitas vezes, a reindexação dos documentos coletados. Frequentemente, estudos nesse formato realizam análises temáticas dos termos indexados não levando em consideração outras partes do texto que podem fornecer temas relevantes. Essa dificuldade esbarra na ausência de uma padronização terminológica das palavras tornando mais simplificada a análise do vocabulário previamente indexado (SANTOS, 2015).

Contudo, não apenas o processo de padronização é considerado como uma problemática na análise temática, mas também, a escolha das ferramentas e do formato mais adequado de visualização da informação é um dos principais desafios a serem vencidos, como afirma Pinto *et al.* (2017, p. 399):

> Entende-se que, uma análise que permita refletir sobre as diversas relações que vão muito além do recorte espacial adotado, torna-se um exercício extremamente importante para que se vislumbre e se produza uma ciência que se preocupa em perceber a realidade de maneira mais totalizante. Entretanto, a elaboração de visualizações gráficas é, porém, extremamente dependente da qualidade dos dados de partida, ou seja, da consistência dos registros dos repositórios de informação. É nessa questão que se estabelece a necessidade de trabalhar a representação da informação conforme explicitado anteriormente.

Portanto, os dados que vão alimentar as ferramentas de visualização precisam estar de acordo com os parâmetros estabelecidos pelo pesquisador, com as variáveis definidas e a

inconsistência eliminada no processo de obtenção e processamento. Assim, os formatos utilizados para a apresentação dos dados não irão sofrer com as fragilidades dos dados brutos.

Com o objetivo de fornecer um método mais adequado, que possibilite a amplitude na análise temática, é importante a adoção de um padrão que permita a realização desta operação, para tanto, a metodologia de "modelagem de tema" pode ser um caminho aceitável na visualização temática, como afirmam os autores:

> Por modelagem de tópicos, é conhecida uma série de técnicas que permitem analisar automaticamente os tópicos abordados nos documentos que compõem uma coleção; Isso inclui a detecção de tais temas, sua intensidade e sua vinculação com outras informações adicionais disponíveis nos documentos; por exemplo, o idioma ou a data (GÁRCIA-MARCO; FIGUEROLA; PINTO, 2020, p. 06, tradução nossa).[4]

De acordo com os autores, esse modelo apresenta a ideia de que num conjunto de documentos existe um conjunto de temas e cada documento representa uma proporção específica de cada tema. Contudo, existe a possibilidade que um só documento não apresente nenhum dos temas centrais encontrados na maioria dos documentos, e de mesma maneira, pode existir um documento que possua mais temáticas encontradas no todo. Os autores supracitados, Gárcia-Marco, Figuerola e Pinto (2020, p. 06, tradução nossa), asseguram:

> A missão da modelagem temática consiste, portanto, em identificar o conjunto de temas do acervo documental e em estabelecer a proporção de cada tema em cada documento. Essas operações são baseadas na coocorrência de palavras nos mesmos documentos; e permitem estabelecer conjuntos de palavras definidoras, com maior ou menor peso, para cada tópico. A presença de uma ou outra palavra em cada documento também permite estimar a porcentagem ou proporção que cada tópico desempenha no conteúdo daquele documento.[5]

Neste sentido, os processos e técnicas que envolvem a análise temática de um conjunto de documentos necessitam de uma delimitação metodológica que possa fundamentar os resultados obtidos nas análises. De forma evolutiva, os elementos fundantes que

[4] Por modelado de temas se conoce una serie de técnicas que permiten analizar de manera automática los temas tratados en los documentos que conforman una colección; esto incluye la detección de dichos temas, su intensidad y su vinculación con otra información adicional disponible sobre los documentos; por ejemplo, el idioma o la fecha. (GÁRCIA-MARCO; FIGUEROLA; PINTO, 2020, p. 06).

[5] La misión del modelado de temas consiste, pues, en identificar el conjunto de temas de la colección documental y en establecer la proporción de cada tema en cada documento. Estas operaciones se basan en la coocurrencia de palabras en los mismos documentos; y permiten establecer conjuntos de palabras definitorias, con mayor o menor peso, de cada tema. La presencia de unas u otras palabras en cada documento permite también estimar el porcentaje o proporción que cada tema juega en el contenido de ese documento. Gárcia-Marco, Figuerola e Pinto (2020, p. 06).

proporcionam a completude das análises propostas devem ser continuamente testados e verificados quanto à sua eficácia.

Uma das possibilidades de visualização capazes de analisar esses elementos é a análise de agrupamento ou de "cluster". Nesse tipo de análise, com a utilização de diferentes algoritmos, é possível detectar divisões entre os grupos de palavras que se baseiam na similaridade e a interconexão entre os termos analisados, Nadzar, Bakri e Ibrahim (2017, p. 93, tradução nossa) conceituam sobre esse tipo de análise:

> Uma vez identificada ou estabelecida a relação ou distância entre as palavras, será gerado o agrupamento de palavras de forma a representar o conjunto de palavras significativas que possuem relação entre uma e outra. A curta distância entre as palavras significa a forte relação e conexão de importância entre si. Em um determinado mapa, poucos clusters são gerados e certamente existem algumas conexões entre os clusters. Portanto, no mapa bibliométrico, são ferramentas úteis para estudar a estrutura e a dinâmica da pesquisa científica.[6]

Isto posto, o distanciamento existente entre as palavras num determinado agrupamento, indica o grau de conexão entre as palavras, e esse tipo de análise pode facilitar a compreensão da dinâmica da pesquisa, e como ela se relaciona tematicamente.

Apesar da pesquisa bibliométrica ser propensa a realizar uma análise quantitativa dos estudos publicados na literatura, é importante ressaltar que os estudos bibliométricos também buscam identificar o comportamento em evolução dos aspectos qualitativos da ciência. Neste sentido, a análise de coocorrência de palavras surge como alternativa e fornece um aspecto direcionado aos comportamentos das publicações científicas.

Essa linha de estudos sobre coocorrência de palavras, também chamada de *co-word*, surgiu como uma corrente da subárea da bibliometria denominada de cocitação. A cocitação pode ser definida como *the frequency with which two items of earlier literature are cited together by the later literature*, Small (1973). Nesse âmbito, a análise de *co-word* vem sendo utilizada para verificar a coocorrência de palavras nas publicações científicas de um assunto específico. Os autores Ding *et al*. (2001, p. 818, tradução nossa) sustentam essa afirmação, quando colocam:

[6] Once the relationship or distance between words were identified or establish, the cluster of words will be generated in order to represent the group of significance words which has the relationship between one and another. The close distance between words means the strong relationship and importance connection between each other. In one certain map, there are few clusters are generated and there are certainly some connections between clusters. Therefore, in bibliometric map, it is useful tools to study the structure and the dynamics of scientific research. Nadzar, Bakri e Ibrahim (2017, p. 93).

> A análise de co-word reduz e projeta os dados em uma representação visual específica com a manutenção das informações essenciais contidas nos dados. Baseia-se na natureza das palavras, que são importantes portadoras de conceitos, ideias e conhecimentos científicos. [7]

Sendo assim, a análise de coocorrências surge como uma técnica importante na visualização de um campo científico quando se atém aos critérios relacionados aos conceitos e ideias das palavras. A descrição semântica de cada termo é algo relevante neste contexto.

Com o objetivo de elucidar quais foram os principais trabalhos identificados que se relacionam tematicamente com à pesquisa aqui desenvolvida, foi elaborada a seção a seguir.

2.1.2 Trabalhos Relacionados a Indicadores Temáticos

Este trecho objetiva apresentar os principais trabalhos e autores relacionados a IT. Desta feita, tem-se subseções que discutem: Cartografia temática e cartografia bibliométrica; e Bibliometria temática e mapeamento bibliométrico.

2.1.2.1 Cartografia Temática e Cartografia Bibliométrica

Nesta subseção são referenciados os trabalhos que possuem relação com cartografia temática ou cartografia bibliométrica no seu escopo, além de uma breve definição sobre o tema em destaque na seção.

De acordo com Ding, Chowdhury e Foo (2001), a cartografia bibliométrica é um método de visualização de cocitação de palavras em um conjunto de documentos. Na análise de cocitação de palavras, ocorre a redução e a projeção dos dados em uma representação visual específica com a manutenção das informações essenciais contidas nos dados. Esse método utiliza a (MDS) Análise Multidimensional Escalar, para criar a visualização dos mapas das matrizes.

> O escalonamento multidimensional (MDS) é um conjunto de técnicas usadas para criar exibições visuais (mapas) a partir de matrizes. A principal saída do MDS é uma exibição de pontos em duas ou três dimensões. Os pontos são colocados no mapa de acordo com sua proximidade na matriz de co-citação do autor (onde valores altos refletem semelhanças altas). Os pontos que representam os autores com semelhanças altas serão colocados próximos uns dos outros, enquanto os pontos que representam os autores com semelhanças

[7] Co-word analysis reduces and projecs the data into a specific visual representation with the maintenance of essential information containing in the data. It is based on the nature of words, wich are the important carrier of scientific concepts, idea and knowledge. Ding et al (2001, p. 818).

> baixas serão colocados mais afastados no mapa. (DING; CHOWDHURY; FOO, 1999, p. 69, tradução nossa).[8]

Dessa maneira, essa técnica, originalmente desenvolvida para representar visualmente a cocitação de autores, foi gradualmente sendo utilizada para outros tipos de análise de cocitação, como a *co-word analysis*. Outros autores afirmam da relevância em se buscar representar graficamente esse tipo de análise.

Segundo Noyons e Van Raan (1994, p. 159, tradução nossa), as vantagens da utilização da representação por cartografia bibliométrica, são:

> Existem várias vantagens importantes no uso de tais representações cartográficas. A visualização de massas complexas de dados oferece uma visão geral mais completa em menos tempo. Além disso, as informações visuais são mais facilmente lembradas. Outro ponto importante é a redução da informação. O mapeamento bibliométrico permite a filtragem de características significativas.[9]

Neste sentido, a representação cartográfica oferece uma série de vantagens que devem ser levadas em consideração no processo de escolha do método de visualização, dentre elas a possibilidade de uma melhor visualização da informação e maior filtragem dos dados. No mesmo texto, os autores afirmam ainda que a utilização de uma abordagem por cartografia facilita a visualização por subcampos semelhantes com agrupamentos de co-palavras. Nesta perspectiva, a construção desses mapas permite uma melhor definição das ligações entre os termos analisados.

Como o objetivo de favorecer a compreensão sob esse tema, apresenta-se o estudo desenvolvido por Ding, Chowdhury e Foo (2001). O artigo "*Bibliometric cartography of information retrieval research by using co-word analysis*" teve o propósito de mapear a estrutura intelectual do campo científico da RI no período de (1987-1997). Os autores buscaram realizar uma análise de coocorrência de palavras com o objetivo de identificar padrões e tendências no campo da RI. Foram utilizadas 3327 palavras de 2012 artigos indexados na *Science Citation Index* (SCI) e *Social Science Citation Index* (SSCI), no período citado. Para a construção desse *corpus* de palavras-chave, os autores fizeram coletas nos

[8] Multidimensional scaling (MDS) is a set of techniques used to create visual displays (maps) from matrices. The major output of MDS is a display of points in two or three dimensions. Points are placed on the map according to their proximity in the author co-citation matrix (where high values reflect high similarities). Points representing authors with high similarities will be placed close together, while points representing authors with low similarities will be placed farther apart in the map (DING; CHOWDHURY; FOO, 1999, p. 69).

[9] There are several important advantages of using such cartographical representations. Visualization of complex masses of data offers a more complete overview in less time. In addition, visual information is more easily remembered. Another important point is the reduction of information. Bibliometric mapping allows the filtering of significant features. Noyons e Van Raan (1994, p. 159)

campos das palavras-chave dos documentos e extraíram, manualmente, palavras-chave dos títulos e resumos com base no LISA *thesaurus*, LCSH e no *Thesaurus of Information Technology Terms*.

Após uma filtragem dos termos com frequência maior que 2, os autores puderam utilizar 193 palavras na análise, para o período analisado. Logo, os autores construíram uma matriz de 193 x 193 palavras e fizeram a correlação dos termos, inserindo a frequência com base na coocorrência. Os principais resultados apontaram 5 grandes grupos de cerca de 50 palavras cada, indicando as relações temáticas mais próximas com a RI.

A segunda pesquisa a ser destacada é a de Kobashi, Díaz e Santana (2016), intitulada "Cartografia temática e de colaboração em organização do conhecimento no Brasil (2000-2010)". Neste trabalho foram utilizados métodos bibliométricos com o objetivo de criar indicadores temáticos sobre o tema "organização da informação"; foram empregados 646 registros distribuídos por 216 artigos de periódicos e 176 de trabalhos de eventos da área da Ciência da Informação.

Os pesquisadores realizaram os seguintes estudos: construíram gráficos apresentando os índices de frequência das palavras-chave utilizadas nos trabalhos; elaboraram um gráfico com os temas mais frequentes nos trabalhos de evento; realizaram um estudo com a evolução da frequência dos termos de trabalhos de eventos; formularam um gráfico de representação hierárquica associando os temas, descritos pelas palavras-chave, relacionando-os às instituições às quais os autores dos artigos e trabalhos estavam vinculados; e, por fim, elaboraram gráficos de redes sociais, associando os pesquisadores aos temas mais trabalhados. Percebe-se, então, após apreciação do estudo supracitado, que ele apresenta uma grande quantidade de resultados e exemplos a respeito da representação temática da informação sobre uma área específica do conhecimento.

Em seguida, descreve-se o estudo desenvolvido por Pinto *et al*. (2017), intitulado "Cartografia temática da produção técnico-científica da Embrapa destinada à agricultura familiar". Nessa pesquisa, os autores objetivaram avaliar a produção intelectual da Embrapa destinada à agricultura familiar, registrada nas publicações técnico-científicas editadas pela Empresa. Foram analisados 65.535 arquivos entre os anos de 2011 e 2016. Para a construção das análises, os autores optaram pelos registros que continham a indexação completa das palavras-chave e do campo "categoria de assunto"; dessa forma, foram definidas 17 categorias relacionadas aos temas. Os autores elaboraram um gráfico interativo [10] contendo as 10

[10] Disponível em: http://bit.ly/2voskAa.

principais categorias adotadas nos documentos e uma figura, representando a cartografia temática, com os temas trabalhados, vinculando-os a todos os estados do Brasil em que foram produzidos os estudos. Os resultados indicaram como as publicações se organizam e se apresentam no espaço.

Logo após, a partir das relações existentes da cartografia temática, propôs-se discutir sobre a área da bibliometria temática e o Mapeamento Bibliométrico, que são áreas do conhecimento que se relacionam diretamente com a proposta desta pesquisa.

2.1.2.2 Bibliometria Temática e Mapeamento Bibliométrico

Nesta subseção são referenciados os trabalhos que possuem relação com bibliometria temática ou mapeamento bibliométrico no seu escopo.

De acordo com Van Eck *et al.* (2010, p. 581), o mapeamento bibliométrico é uma ferramenta poderosa para estudar e analisar a dinâmica dos campos científicos. Os autores afirmam ainda que os pesquisadores podem utilizar essa ferramenta para compreender melhor determinado campo científico, contudo, tal tarefa não é trivial.

> Os mapas de termos são semelhantes aos mapas de palavras compartilhadas, exceto que podem conter qualquer tipo de termo em vez de apenas termos de uma única palavra ou apenas palavras-chave. Ao construir um mapa bibliométrico, primeiro deve-se selecionar os objetos a serem incluídos no mapa. No caso de um mapa que contém autores ou periódicos, isso geralmente é bastante fácil. Para selecionar os autores ou periódicos importantes em um campo, geralmente se pode simplesmente confiar na contagem de citações. No caso de um mapa de termos, as coisas não são tão fáceis. Na maioria dos casos, é muito difícil selecionar os termos importantes em um campo. A seleção de termos com base em sua frequência de ocorrência em um *corpus* de documentos normalmente produz muitas palavras e frases com pouco ou nenhum significado específico de domínio. A inclusão de tais palavras e frases em um mapa de termos é altamente indesejável por dois motivos. Primeiro, essas palavras e frases desviam a atenção do que é realmente importante no mapa. Em segundo lugar, e ainda mais problemático, essas palavras e frases podem distorcer toda a estrutura mostrada no mapa. Como não há uma maneira fácil de selecionar os termos a serem incluídos em um mapa de termos, a seleção de termos geralmente é feita manualmente com base na opinião de especialistas (VAN ECK et al, 2010, p. 582, tradução nossa).[11]

[11] Term maps are similar to co-word maps except that they may contain any type of term instead of only single-word terms or only keywords. When constructing a bibliometric map, one-first has to select the objects to be included in the map. In the case of a map that contains authors or journals, this is usually fairly easy. To select the important authors or journals in a field, one can usually simply rely on citation counts. In the case of a term map, things are not so easy. In most cases, it is quite difficult to select the important terms in a field. Selection of terms based on their frequency of occurrence in a *corpus* of documents typically yields many words and phrases with little or no domain-specific meaning. Inclusion of such words and phrases in a term map is highly

Os mapas de termos são mapas que apresentam um conjunto de palavras. Para a construção desse mapa é necessária a seleção correta dos elementos que serão analisados. A seleção dos termos mais importantes dentro de um campo exige uma quantificação das palavras mais frequentes, contudo, esse procedimento produz palavras com pouco significado. Com o objetivo de otimizar esse procedimento de escolha das palavras mais relevantes, a identificação ou indexação automática dos termos num determinado *corpus* de documentos é um caminho interessante e que pode facilitar a visualização do domínio analisado. (VAN ECK *et al*., 2010)

Neste sentido, evidencia-se o artigo elaborado por Noyons e Van Raan (1994), intitulado "*Bibliometric cartography of scientific and technological developments of an R & D field*". Neste artigo, os autores investigaram o mapeamento bibliométrico como uma ferramenta analítica para o estudo dos aspectos importantes relacionados à Ciência e à Tecnologia. Os pesquisadores desenvolveram a construção de mapas da ciência baseados na coocorrência de palavras-chave de publicações e patentes. A área estudada foi a Optomecatrônica e suas relações. A comparação entre as duas áreas (Ciência e Tecnologia) permitiu identificar as interações que modificaram os diferentes subcampos. Logo, foi possível averiguar as possíveis ligações existentes entre a Ciência e a Tecnologia na área da Optomecatrônica.

Posteriormente, expõe-se o estudo elaborado pelos pesquisadores Van Eck e Waltman (2009), no artigo intitulado "*Software survey: VOSviewer, a computer program for bibliometric mapping*", os autores apresentaram um software voltado para representação gráfica de mapas bibliométricos que permite construir mapas de palavras-chave com base em dados de coocorrência de palavras. O *software* possui a capacidade de adotar as técnicas de visualização do escalonamento multidimensional, a mais comum no campo da bibliometria, para a representação gráfica de grandes mapas bibliométricos. No caso do escalonamento, o tamanho da distância entre duas palavras no texto é o que determina a relação de força entre os dois elementos. Neste caso, quanto menor a distância mais forte é a relação. Assim, o *software* consegue exibir um grande conjunto de dados com pouca interferência ou "poluição visual".

---

undesirable for two reasons. First, these words and phrases divert attention from what is really important in the map. Second and even more problematic, these words and phrases may distort the entire structure shown in the map. Because there is no easy way to select the terms to be included in a term map, term selection is usually done manually based on expert judgment (VAN ECK et al, 2010, p. 582).

A posteriori, relata-se a pesquisa intitulada "*A Comparison of Two Techniques for Bibliometric Mapping: Multidimensional Scaling and VOS*", desenvolvido por Van Eck *et al.* (2010). O trabalho apresenta a análise comparativa entre dois tipos de técnicas utilizadas no mapeamento bibliométrico: a técnica de escalonamento multidimensional (MDS) e a técnica de visualização de semelhanças (VOS). Os autores afirmam que o objetivo das duas técnicas é o mesmo, o de analisar a proximidade e a distância entre os itens de um documento.

No caso da MDS, ela propõe medir essa relação pelo cálculo da semelhança entre os itens, dessa forma, a distância entre dois itens reflete o grau de semelhança ou relação existente. Portanto, quanto mais forte for a relação entre os itens, menor a distância entre eles. Já, no caso da técnica VOS, a ideia é identificar a soma ponderada das distâncias quadradas entre todos os pares de itens identificados, assim, o quadrado da distância entre um par de itens é ponderado pela semelhança entre os itens.

Ao final, nas conclusões, os autores afirmam que a técnica VOS fornece uma representação gráfica mais satisfatória do conjunto de dados analisados. A técnica MDS apresenta um certo índice tendencioso na representação, ao localizar itens mais importantes no centro e menos importantes na periferia. A VOS não possui essa característica além de produzir melhores resultados para um conjunto de dados considerado médio e grande (VAN ECK *et al.*, 2010, p. 2414).

Em seguida, expõe-se a pesquisa desenvolvida por Heersmink *et al.* (2011). Neste trabalho, os autores buscaram apresentar uma análise da área de informática e informação ética por meio de um mapeamento bibliométrico. Eles apresentaram as relações presentes em 400 termos-chave no campo por meio das relações entre os conceitos e os principais tópicos utilizados. A metodologia consistiu na análise dos títulos e resumos de mil artigos publicados em 12 principais revistas e 3 conferências no campo da computação e informação ética. Os autores levaram em consideração, que os títulos e resumos poderiam representar o conteúdo completo do artigo, não prejudicando, consequentemente, a análise temática proposta.

Neste estudo, foram utilizados 1027 artigos publicados de 2003 a 2009 nas revistas e conferências. Com o objetivo de realizar um mapeamento bibliométrico e produzir as representações visuais necessárias para identificar as relações das palavras coocorrentes nos títulos e resumos, os autores selecionaram os 400 sintagmas nominais mais relevantes do *corpus* considerado. Com o intuito de realizar as representações visuais das coocorrências de palavras analisadas, foi utilizado o software VOSviewer[12]. O *software* construiu um mapa de

[12] (VAN ECK; WALTMAN, 2010). Disponível em: www.vosviewer.com. Acesso em: 16 out. 2021.

termos baseados nas frequências de coocorrências de palavras, assim, foi possível identificar a proximidade entre os termos e as possíveis relações entre eles.

Adiante, evidencia-se o estudo "*A bibliometric mapping of the structure of STEM education using co-word analysis*", realizado por Assefa e Rorissa (2013). As autoras buscaram analisar um conjunto de campos da educação, das seguintes áreas: Matemática, Engenharia, Ciências Ambientais, Ciências da Vida, Agricultura, e Ciências da Computação. Para facilitar a identificação dessas áreas, as autoras utilizaram a sigla "STEM" e STEM *education* para as análises propostas. Para isso, utilizaram um *corpus* textual de 7.265 documentos e analisaram os campos dos títulos, palavras-chave, descritores e resumos desses documentos. Esses arquivos incluíam livros, anais de conferências, artigos de periódicos, dissertações, teses e relatórios, publicados de 1901 a 2010.

Para a análise dos resultados, as pesquisadoras recorreram ao pacote de software de mapeamento do T-LAB[13]. Com esses sistemas, elas puderam verificar a coocorrências das palavras por meio do agrupamento dos lemas (raiz lexical) de cada palavra. Então, elas conseguiram identificar as principais áreas do conhecimento que se relacionam com a "STEM *education*".

Há também um estudo desenvolvido e publicado por Santos (2015), intitulado "Organização e representação do conhecimento: bibliometria temática em artigos de periódicos brasileiros". Nessa pesquisa, a autora realizou uma investigação exploratória, em cerca de 150 artigos indexados na BRAPCI[14], sobre Organização da Informação e temas afins, publicados entre 1996 e 2013. Ela adotou o software NVivo[15] para gerar as análises de agrupamento e identificar as similaridades semânticas entre os termos. Como resultado, foram gerados 2 diagramas de nós por similaridade, um de palavras e outro de codificação. A pesquisa identificou, ainda, correlações temáticas na área de Organização e Representação do Conhecimento.

Posteriormente, relata-se o estudo realizado por Nadzar, Bakri e Ibrahim (2017). Nesse trabalho, os autores buscaram identificar o progresso da pesquisa que foi alcançado pelos pesquisadores na Malásia, analisando a base de dados Scopus por meio de um mapeamento bibliométrico. Para isso, coletaram-se 2360 artigos publicados no período entre 1995 e 2015 com o objetivo de analisar a coocorrência das palavras nos títulos e resumos. Neste ínterim, os autores encontraram agrupamentos temáticos em torno das seguintes áreas: Medicina,

[13] Disponível em: www.tlab.it/en/.
[14] Disponível em: https://brapci.inf.br/index.php/res/.
[15] Disponível em: https://www.software-shop.com/producto/nvivo_portugues.

Farmácia, Agricultura, Meio Ambiente, Silvicultura, Saúde do Adulto e Saúde nos Países em Desenvolvimento. Além de terem sido consideradas as análises de coocorrência também foram verificados os termos mais frequentes ao longo de 20 anos (NADZAR; BAKRI; IBRAHIM, 2017).

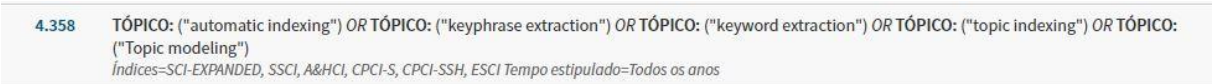
Em seguida, com o propósito de apresentar os resultados sobre o tema IA relacionando-os ao campo do Mapeamento bibliométrico, foi elaborada uma seção que pudesse relacionar tematicamente ambas as áreas, o que pode ser verificado a seguir.

2.1.3 Indexação Automática aplicada ao Mapeamento bibliométrico

Com o objetivo de identificar trabalhos que se aproximem da pesquisa aqui desenvolvida, e que utilizaram a IA juntamente com o mapeamento bibliométrico, foi necessário realizar uma busca numa base de dados científica internacional que pudesse fornecer conteúdos relacionados à proposta de pesquisa. Isto posto, foi realizada a seguinte busca na *WoS*[16]:

1ª Etapa – Foi inserida a seguinte expressão de busca por tópicos dentro da base, utilizando a busca básica: ("automatic indexing" OR "keyword extraction" OR "keyphrase extraction" OR "topic indexing" OR "Topic modeling"). A busca se deu em todas as bases indexadas dentro da WoS e por todo o período estipulado. O resultado obtido foi de 4.358 documentos da principal coleção da *WoS* (figura 2).

Figura 2 – Descrição da Busca 1

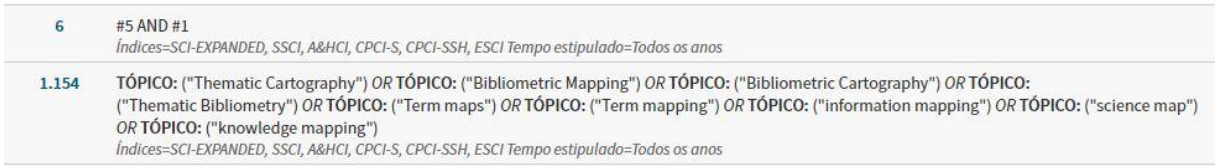


Fonte: *Web of Knowledge* (2021).

2ª Etapa – Nesta segunda etapa foi inserida a seguinte expressão de busca por tópicos dentro da base, utilizando a busca básica: ("Thematic Cartography" OR "Bibliometric Cartography" OR "Thematic Bibliometry" OR "Bibliometric Mapping" OR "Term maps" OR "Term mapping" OR "information mapping" OR "science map" OR "knowledge mapping"). Os resultados obtidos foram de 1.154 documentos da principal coleção da *WoS* (figura 3).

[16] Busca realizada no dia 29/01/2021 às 16:40h.

Figura 3 – Descrição da Busca 2



Fonte: *Web of Knowledge* (2021).

3ª Etapa – Com as expressões de busca salvas no perfil, foi realizada uma combinação das duas buscas anteriores usando o operador lógico "AND". Enquanto, na opção de busca avançada, foi possível combinar as expressões de busca e obter a intersecção de trabalhos que tratam das duas temáticas de forma simultânea. O resultado obtido foi de seis documentos que tratam do perfil selecionado, que serão analisados a seguir:

> Mapping human resource development: Visualizing the past, bridging the gaps, and moving toward the future. Source: HUMAN RESOURCE DEVELOPMENT QUARTERLY. Abstract: Using topic mapping techniques, we provide a review of the 3,236 articles published in the five premier HRD journals between 1990 and 2019. We map the key terms evidencing the emergence of five major topic clusters within HRD scholarship: nature and identity of HRD, HRD interventions and outcomes, national HRD, career development, and HRD in academia. Nature and identity of HRD depicts a movement from establishing foundations to expanding horizons. HRD interventions and outcomes reflects a spectrum of research interests from exploring processes to examining desired outcomes. National HRD shows that research topics are increasingly moving toward international and global settings. Career development demonstrates a shift from emphasizing organizational careers to advocating for under-represented career actors. HRD in academia encompasses topics related to educating HRD professionals and supporting professional development in other fields. We provide a synthesis of the clusters, create a taxonomy of topic areas, and identify the mature, nascent, trending, and growing concepts in HRD. In doing so, our paper provides an overview of where we are in HRD scholarship. We then suggest collaborative, competitive, and configurational boundary work as strategies that HRD scholars can use purposefully to influence future HRD research directions. Our findings inform HRD researchers and can guide interested practitioners in their search for actionable knowledge in HRD (SHIRMOHAMMADI *et al.*, 2020, p. 197).

No primeiro artigo, os autores realizaram uma revisão em 3236 artigos sobre a área de Desenvolvimento de Recursos Humanos, publicados entre 1990 e 2019. Eles mapearam os termos mais importantes que destacavam cinco agrupamentos temáticos. Para a elaboração dos gráficos nas análises, os pesquisadores utilizaram o software Vosviewer de Van Eck e Waltman (2011). Assim sendo, eles obtiveram os gráficos de correlação temática e os mapas de calor identificando os agrupamentos de palavras.

> The evolution of social health research topics: A data-driven analysis. Source: SOCIAL SCIENCE & MEDICINE. Abstract: The realm of social health has not yet been properly established in terms of fixed definitions, concepts, and research areas. This study attempts to define social health using macro and micro perspectives and explores trends in social health research by mapping their topics and fields. We used Latent Dirichlet allocation (LDA) topic modeling, which allows the extraction of key terms and topics derived from a large volume of literature. We traced the evolution of research topics from past (the literature that "present" articles cited), present (existing journal articles on social health), to future (the literature which cited the articles) studies based on connections between citations. The datasets were collected by the query terms "social health" in the Scopus database, including title, abstract, and keywords of journal articles. We collected a total of 443 articles from recent social health literature, 6588 articles from past literature that the recent articles on social health cited, and 2680 articles from future literature in which recent social health articles were cited. We defined social health as positive interaction that increases individual engagement in social life at the micro level, and the high degree of social integration that deals with collective problems in society at the macro level. The results of LDA showed that social health research has developed into seven fields: Health Care Delivery; Vulnerable Groups; Measurement; Health Inequality; Social Network and Empowerment; Clinical/Physical Health; and Mental/Behavioral Health. Based on citation relationships, topics grounded in an individual/micro perspective have grown increasingly specialized and productive, while topics grounded in a social/macro perspective have stagnated or was underexplored. Our findings imply that social health studies should follow a more interdisciplinary approach to integrate current health models of individual-centered treatments with social science concerns on building collective capacity for social well-being (CHO; PARK; SONG, 2020, p. 01).

No segundo artigo, os autores realizaram uma coleta em 443 artigos recentes sobre a área de Saúde Social, 6588 artigos da literatura passada, que foi citada pelos 443 artigos, e 2680 artigos da literatura futura da área de Saúde Social, que vão ser citados. Os autores utilizaram o método de MT usando modelagem de tópicos com o algoritmo LDA. Esse método permitiu examinar as relações entre os termos e extrair tópicos de suas estruturas. A implementação dos métodos propostos foi disponibilizada no site do Github[17]

> When Public Health Research Meets Social Media: Knowledge Mapping From 2000 to 2018. Source: JOURNAL OF MEDICAL INTERNET RESEARCH. Abstract: Social media has substantially changed how people confront health issues. However, a comprehensive understanding of how social media has altered the foci and methods in public health research remains lacking. This study aims to examine research themes, the role of social media, and research methods in social media-based public health research published from 2000 to 2018. A dataset of 3419 valid studies was developed by searching a list of relevant keywords in the Web of Science and PubMed databases. In addition, this study employs an unsupervised text-mining technique and topic modeling to extract research themes of the

[17] Disponível em: https://github.com/Yonsei-TSMM/social_health.

> published studies. Moreover, the role of social media and research methods adopted in those studies were analyzed. This study identifies 25 research themes, covering different diseases, various population groups, physical and mental health, and other significant issues. Social media assumes two major roles in public health research: produce substantial research interest for public health research and furnish a research context for public health research. Social media provides substantial research interest for public health research when used for health intervention, human-computer interaction, as a platform of social influence, and for disease surveillance, risk assessment, or prevention. Social media acts as a research context for public health research when it is mere reference, used as a platform to recruit participants, and as a platform for data collection. While both qualitative and quantitative methods are frequently used in this emerging area, cutting edge computational methods play a marginal role. Social media enables scholars to study new phenomena and propose new research questions in public health research. Meanwhile, the methodological potential of social media in public health research needs to be further explored (ZHANG et al, 2020, p. 84).

O terceiro artigo trata de um estudo que teve como objetivo examinar temas e métodos de pesquisa, sobre o papel das mídias sociais, examinando pesquisas de Saúde Pública baseadas em mídias sociais publicadas de 2000 a 2018. O estudo analisou os metadados de um conjunto de 3419 artigos validados. Os campos analisados foram: título, autores, título do periódico, resumo, palavras-chave e referências citadas. Os pesquisadores adotaram a técnica de MT não supervisionada de modelagem de tópicos (LDA), para extrair temas de pesquisa dos estudos publicados. Além disso, foi analisado o papel das mídias sociais e os métodos de pesquisa adotados nesses estudos. Os resultados indicaram 25 temas de pesquisa, cobrindo diferentes doenças, vários grupos populacionais, sobre saúde física e mental e outras questões significativas.

> Dynamics of topic formation and quantitative analysis of hot trends in physical science. Source: SCIENTOMETRICS. Abstract: Successful research in the face of increasing complexity of modern scientific knowledge together with diversity and depth of the studied problems requires an understanding of the structure and evolution of trends in science. Available digital records open wide possibilities for statistical analysis of scientific publications and related metadata for topic modeling and evolution, knowledge mapping, citation indexing, etc. We investigate dynamical properties of the physical topics using analysis of temporal evolution of proximity measure for word pairs related to the mutual information. We use full-text conceptualization of content of scientific documents provided by the ScienceWISE platform for topic mapping, trend analysis and detection of hot topics together with relevant papers retrieval. We found that time evolution of relative mutual information distance reveals a hidden topic structure and could be used for quantitative analysis of current trends in scientific research (CHUMACHENKO *et al.*, 2020, p. 87).

A quarta pesquisa trata de uma análise dos conceitos sobre Física de Alta Energia. Para isso, utilizou-se a ontologia *science-wise* (SW) e de informações sobre as frequências de

citações de conceitos no texto completo dos documentos científicos (incluindo o resumo do título e o texto principal), relacionados para esse campo da Física. O sistema *science-wise* (SW) surgiu como resultado de uma colaboração entre físicos e cientistas da computação da EPFL e CERN. O SW usa os métodos modernos de recuperação de informações, análise de texto e análise de dados estatísticos para processar grandes quantidade de publicações e oferecer um sistema de recomendação semântica. Dessa forma, os autores recuperaram os conceitos de Física de Alta Energia, dentro da coleção. Foram utilizados 451.523 documentos publicados de 1988 a 2018 no ArXiV: HEP *e-print server*. O *corpus* gerou uma ontologia composta de 16.370 conceitos fornecidos pela *science-wise*. Após o processamento, foi possível analisar e calcular a semelhança entre os conceitos adotando o uso de uma métrica com base em informações mútuas normalizadas (NMI).

> A data-driven analysis of the knowledge structure of library science with full-text journal articles. Source: JOURNAL OF LIBRARIANSHIP AND INFORMATION SCIENCE. Abstract: In previous studies, full-text analyses and mining techniques have not been combined to identify and trace changes in the knowledge trends of library science over the past 20 years (1997-2016). Thus, to grasp the knowledge trends of library science at a fine-grained level, this study analyzes full-text journal articles from six top-ranked library science journals by applying text-mining techniques such as co-word analysis, text summarization, and topic modeling. Visualization tools were used to map the knowledge structure of library science. The findings indicate that, during the past 20 years, library science has developed into an interdisciplinary knowledge structure that integrates librarianship topics with a range of other fields, generating major topics that include the academic library, the digital library, research methodology, library marketing, information retrieval, digital information, document citation, and so on. In the past ten years, the library science discipline has focused increasingly on research methodology and evaluation and become more concerned with digital information management (TIMAKUM; KIM; SONG, 2020, p. 345).

No quinto estudo analisado, os autores buscaram compreender as tendências de conhecimento da Biblioteconomia em um nível aprofundado. Para isso, a pesquisa analisou 4023 artigos publicados em seis periódicos de Biblioteconomia com melhor índice de classificação, no período de 1997 a 2016. Destarte, foram aplicadas técnicas de MT, como análise de co-palavras, resumo de texto e modelagem de tópicos. Como método, adotou a ferramenta de código aberto 'yTextMiner'[18] ela é uma ferramenta de MT desenvolvido em JAVA. Além disso, utilizou o Gephi[19] e Vosviewer[20] como *softwares* para visualização da topologia tópica da área. Os principais resultados apontaram que num período de 10 anos

[18] Disponível em: http: //informatics.yonsei.ac.kr:8080/yTextMiner/home.html.
[19] Disponível em: https://gephi.org/.
[20] Disponível em: http://www.vosviewer.com/.

(2007-2016), ocorreram diversas concentrações temáticas em áreas como: informação digital e metodologia de gestão e pesquisa. O estudo também demonstrou que as técnicas de MT são uma forma de analisar e extrair conhecimento das estruturas disciplinares de grandes conjuntos de dados. A pesquisa fornece uma visão atualizada da estrutura de conhecimento do campo da biblioteconomia, descreve a origem dos subcampos constituintes, e auxilia os pesquisadores a identificarem os componentes gerais do campo científico analisado.

> Text mining techniques for patent analysis. Fonte: INFORMATION PROCESSING & MANAGEMENT. Abstract: Patent documents contain important research results. However, they are lengthy and rich in technical terminology such that it takes a lot of human efforts for analyses. Automatic tools for assisting patent engineers or decision makers in patent analysis are in great demand. This paper describes a series of text mining techniques that conforms to the analytical process used by patent analysts. These techniques include text segmentation, summary extraction, feature selection, term association, cluster generation, topic identification, and information mapping. The issues of efficiency and effectiveness are considered in the design of these techniques. Some important features of the proposed methodology include a rigorous approach to verify the usefulness of segment extracts as the document surrogates, a *corpus*- and dictionary-free algorithm for keyphrase extraction, an efficient co-word analysis method that can be applied to large volume of patents, and an automatic procedure to create generic cluster titles for ease of result interpretation. Evaluation of these techniques was conducted. The results confirm that the machine-generated summaries do preserve more important content words than some other sections for classification. To demonstrate the feasibility, the proposed methodology was applied to a realworld patent set for domain analysis and mapping, which shows that our approach is more effective than existing classification systems. The attempt in this paper to automate the whole process not only helps create final patent maps for topic analyses, but also facilitates or improves other patent analysis tasks such as patent classification, organization, knowledge sharing, and prior art searches (TSENG; LIN; LIN, 2007, p. 1216).

No sexto artigo, os autores descreveram uma série de técnicas de MT em documentos de patentes. As técnicas relatadas incluem: segmentação de textos, sumarização, seleção de campos, associação de termos, geração de agrupamentos de palavras, identificação de tópicos e mapeamento de informação. Os resultados confirmam que os resumos gerados por máquina preservam palavras de conteúdo mais importantes do que algumas outras seções para classificação. Para demonstrar a viabilidade, a metodologia proposta foi aplicada a um conjunto de patentes do mundo real para análise e mapeamento de domínio, o que mostra que nossa abordagem é mais eficaz do que os sistemas de classificação existentes. A tentativa neste artigo de automatizar todo o processo não apenas ajuda a criar mapas finais de patentes para análises de tópicos, mas também facilita ou melhora outras tarefas de análise de patentes,

como classificação de patentes, organização, compartilhamento de conhecimento e pesquisas de arte anterior.

Posteriormente, com o objetivo de identificar trabalhos que se aproximem da análise aqui desenvolvida, e que utilizaram a Indexação Automática juntamente com o Mapeamento bibliométrico, foi necessário realizar uma nova busca na *WoS* que pudesse fornecer conteúdos relacionados à proposta de pesquisa. Por conseguinte, foi realizada a seguinte busca na *Web of Science* no dia 22/02/2021 às 14:30h:

1ª Etapa – Foi inserida a seguinte expressão de busca por tópicos dentro da base, utilizando a busca básica: ("automatic indexing" OR "keyword extraction" OR "keyphrase extraction" OR "topic indexing" OR "Topic modeling"). A busca se deu em todas as bases indexadas dentro da WoS e por todo o período estipulado. O resultado obtido foi de 4.384 documentos da principal coleção da WoS.

2ª Etapa – Nesta segunda etapa foi inserida a seguinte expressão de busca por tópicos dentro da base, utilizando a busca básica: ("Thematic Cartography" OR "Bibliometric Cartography" OR "Thematic Bibliometry" OR "Bibliometric Mapping" OR "Term maps" OR

"Term mapping" OR "information mapping" OR "science map" OR "knowledge mapping" OR "Bibliometric" OR "Bibliometry"). Os resultados obtidos foram de 14.453 documentos da principal coleção da WoS.

3ª Etapa – Com as expressões de busca salvas no perfil, foi realizada uma combinação das 2 buscas anteriores usando o operador lógico "AND". Deste modo, na opção de busca avançada, foi possível combinar as expressões de busca e obter a intersecção de trabalhos que tratam das 2 temáticas de forma simultânea. Após o refinamento dos resultados, selecionando artigos de periódicos possuindo tema relacionado à CI, o resultado total foi de 53 documentos, contudo, foram escolhidos 19 trabalhos relevantes, após serem descartadas as ambiguidades e os trabalhos que não possuem relação com o objetivo da busca, os artigos foram descritos no quadro 6:

Quadro 6 – 19 trabalhos relevantes identificados

CITAÇÃO	CARACTERIZAÇÃO DOS ARTIGOS ANALISADOS	TÉCNICA ADOTADA NO ARTIGO
1-(DAUD *et al*., 2021, p. 798)	Produção acadêmica de artigos de 10 principais pesquisadores do ano de 2005 a 2009.	Propõe a implantação do método ***Hot Topics Rising Star Rank*** (**HTRS-Rank**) para encontrar pesquisadores em ascensão por meio da detecção de tópicos mais relevantes.
2-(YEO; JEONG, 2020, p. 2075)	Foram analisados 6.581 documentos publicados durante o período de 2009-2018.	Foram utilizados métodos bibliométricos, que foram empregados da seguinte forma: adoção de análises estatísticas dos resultados de publicações; utilização da modelagem de tópicos baseada em resumos de publicações com séries temporais; utilização da análise de redes sociais de coautorias para verificação da relação existente entre os autores.
3-(KIM; PARK; LEE, 2020, p. 113401)	Metodologia utilizada para realizar uma análise de tendência anual das pesquisas sobre *blockchain* (serviço explorador de criptomoeda), em 231 resumos de artigos relacionados ao tema, publicados no período de 2015 a 2019.	Utilizou o método de modelagem de tópicos denominado Análise Semântica Latente (**LSA**), baseada na metodologia Word2vec (**W2V-LSA**), utilizada para melhor capturar e representar o contexto de um *corpus*.
4-(CHEN; XIE, 2020, p. 1097)	O estudo apresentou uma revisão bibliométrica de análise de sentimento com base em um método de modelagem de tópicos estruturais para obter uma visão ampla do campo de pesquisa.	Foram utilizados métodos como análise de regressão, visualização geográfica, análise de rede social e teste de tendência como principais técnicas bibliométricas.
5-(CHOI; SEO, 2020, p. 592)	Utilizou um total de 426 artigos publicados entre 2000 e 2018 que foram recuperados da *Web do Clarivate Analytics of Science.* Foram utilizadas técnicas bibliométricas para obter os principais resultados, como: análise de rede de colaboração científica entre países, análise de co-ocorrência de palavras, estrutura conceitual e seis tópicos latentes identificados são discutidos.	Os procedimentos utilizaram análises bibliométricas, como: análise de rede de colaboração; análise de co-ocorrência de palavras e tópicos mais relevantes. Para isso, adotou a técnica de modelagem de tópicos de Alocação Latente de Dirichlet (**LDA**) para a obtenção dos dados observados.
6-(GARCIA-MARCO; FIGUEROLA; PINTO, 2020, p. 1)	O estudo contemplou o período de 1978 a 2019. Para obtenção dos temas, os títulos e resumos	Foi realizada uma avaliação da evolução temática da pesquisa em Biblioteconomia e Ciência

	foram tratados pelo método de Alocação de Latente de *Dirichlet* (**LDA**), com o intuito de realizar uma modelagem temática estatística.	da Informação, na língua espanhola, na base de dados LISA usando alocação latente de Dirichlet (**LDA**).
7-(CHEN *et al.*, 2020, p. 0231192)	A pesquisa objetivou analisar texto não estruturado, com o propósito de identificar publicações sobre o cérebro humano e Inteligência Artificial. O período escolhido foi a última década.	Realizou uma modelagem de tópicos estruturais (**MTS**), com adoção de técnicas bibliométricas, para identificar tópicos de pesquisa proeminentes. Foram identificadas as principais correlações temáticas e agrupamentos de tópicos como resultados.
8-(RANAEL *et al.*, 2020, p. 215)	Caracteriza-se pela adoção de três metodologias distintas que auxiliaram na identificação temática das áreas analisadas. Foram identificadas as frequências dos termos, tamanho e origem das comunidades temáticas e ligações entre os tópicos de pesquisa.	Os autores utilizaram os seguintes procedimentos metodológicos para obtenção dos resultados: técnica de contagem de termos; escore de emergência (**EScore**); e a Alocação Latente de Dirichlet (**LDA**).
9-(SHIN *et al.*, 2018, p. 3522)	Objetivou apresentar uma revisão da literatura dos principais artigos de periódicos na área de estudos marítimos publicados entre 1993 e 2017 usando uma técnica de modelagem de tópicos.	Os autores realizaram um mapeamento por meio da alocação Latente de Dirichlet (**LDA**), técnica utilizada para descoberta de dados e relacionamentos entre os documentos de texto.
10-(SILVA *et al.*, 2018, p. 245)	Objetivou analisar o universo de currículos de 6.060 pesquisadores ativos na Filosofia. Para tanto, foram observados 43.345 artigos publicados entre 2007 e 2016, destes, 1.657 indexados na WoS.	Os dados foram extraídos com software ScriptLattes, baseado na linguagem de programação R. Ele extraiu o conjunto de metadados dos currículos dos pesquisadores numa base de dados de pesquisadores chamada (Plataforma Lattes). Para a geração dos gráficos de coautoria foi utilizado o sotware Gephi.
11-(HU *et al.*, 2018, p. 1031)	Adotou-se um processo de medição de similaridade semântica, e um cálculo estatístico para calcular as frequências de "unidades semânticas" em vez de frequências de palavras-chave.	Foi utilizado o método de modelagem de tópicos denominado **Word2vec** (**W2V-LSA**), utilizada para realizar a distribuição de palavras para representar os significados semânticos das palavras-chave.
12-(FIGUEROLA; MARCO; PINTO, 2017, p. 1507)	Os autores utilizaram as referências bibliográficas (título e resumo) da produção acadêmica sobre *Library Information Science* na base de	Foi utilizada a técnica estatística de modelagem de tópicos, chamada de Alocação Latente de Dirichlet (**LDA**), a fim de identificar os principais

	dados LISA no período 1978-2014, que abrangeram 92.705 documentos.	tópicos e categorias do *corpus* de documentos analisados.
13-(RANAEI; SUOMINEN, 2017, p. 1)	O universo de dados extraídos obteve cerca de 711.296 resumos de patentes concedidas nos Estados Unidos, entre os anos de 1980 e 2014, resultantes de uma pesquisa por "veículo", criando um conjunto de dados complexo de tecnologias de aplicações automotivas.	Utilizou a Alocação Latente de Dirichlet (**LDA**) e a modelagem dinâmica de tópicos, apresentando diferentes padrões temáticos existentes.
14-(HEO *et al*., 2017, p. 45)	Para o *corpus* de análise foram utilizados periódicos e frases-chave. Para tanto, utilizou a base de dados **PubMed** para coletar 46 periódicos de bioinformática do banco de dados MEDLINE. A pesquisa extraiu temas de séries temporais em quatro períodos de 1996 a 2015 para examinar mais a fundo a natureza interdisciplinar sobre o tema "bioinformática".	Utiliza o método de Autor-Conferência-Tópico (**ACT**) para estudar o campo da bioinformática da perspectiva de frases-chave, autores e periódicos. O modelo **ACT** é capaz de incorporar o artigo, o autor e a conferência na distribuição de tópicos simultaneamente
15-(JIANG; QIANG; LIN, 2016, p. 693)	Os autores buscaram analisar 8.280 artigos de pesquisa chineses altamente relacionados ao tema TGP (Projeto Três Gargantas). Para isso foi estabelecido o período de 2001 a 2013. Um modelo de 18 tópicos foi utilizado para descrever a estrutura intelectual de busca.	Utilizou a metodologia adotada para modelagem de tópicos, Alocação Latente de *Dirichet* (**LDA**).
16-(SONG; HEO; LEE, 2015, p. 905)	A técnica consistiu numa análise de produtividade (ano, periódico, autor, termos do *Medical Subject Heading*), para isso, foram utilizadas as técnicas bibliométricas como: análise de rede, frequência de co-ocorrência, centralidade e comunidade, e análise de conteúdo. Para tanto, coletaram-se metadados de 96.081 artigos recuperados do PubMed.	Buscou-se analisar a tendência de tópicos de séries temporais usando a técnica de modelagem de tópicos de regressão multinomial de *Dirichlet*, ou, Alocação Latente de *Dirichet* (**LDA**). Além disso, foram utilizadas as técnicas bibliométricas como: análise de rede, frequência de co-ocorrência, centralidade e comunidade, e análise de conteúdo.
17-(OH; LEE, 2014, p. 4467)	Os autores propuseram a necessidade de novos domínios de pesquisa, e, concluíram que o estudo é significativo, no que diz respeito à análise bibliométrica com resumos e	Foram adotadas três técnicas de mineração de texto: a análise do algoritmo de extração de palavras-chave (**KEA**), a análise de coocorrência e a análise de citações. Além disso

	dados de citações na área de "telecomunicações", bem como no desenvolvimento de softwares que possuem funções de *web services* e técnicas de *text mining*.	os autores utilizaram o software R para a visualização das frequências do termo e a rede de coocorrência entre as publicações.
18-(JEONG; SONG, 2014, p. 776).	Foi proposto um método de análise temporal para utilizar recursos heterogêneos como artigos, patentes e notícias da web de forma integrada. Investigaram-se o fenômeno do intervalo de tempo entre três recursos e de duas áreas acadêmicas, por meio da realização de uma análise de conteúdo baseada em mineração de textos.	Utilizou a técnica de modelagem de tópicos, Alocação Latente de *Dirichet* (**LDA**) para analisar um conjunto de 3 tipos de termos (artigos, patentes e artigos de notícias da web), em dois domínios de pesquisa.
19-(PULGARÍN; GIL-LEIVA, p. 365, 2004)	Adotou um estudo bibliométrico de um *corpus* de 839 referências bibliográficas sobre o descritor "indexação automática", percorrendo o período que vai de 1956 ao ano 2000. Os documentos foram obtidos nas bases da Ciência da Informação, *Library and Information Science Abstracts and Information Science Abstracts.*	Foram realizadas análises bibliométricas por meio da extração de referências bibliográficas de cada artigo e livro analisado. Foram identificados o grau de distribuição dos autores e obras, a obsolescência da literatura e sua dispersão, além da distribuição por tópico ano e tipo de fonte.

Fonte: dados da pesquisa (2021).

Ao analisar-se o quadro, observam-se que 9 artigos utilizaram como técnica a LDA, destacando-se em relação às outras técnicas observadas, que foram utilizadas em uma ou no máximo duas pesquisas. Deste modo, percebe-se que a Alocação Latente de Dirichlet surge como principal técnica para obtenção de agrupamentos temáticos com base num conjunto de documentos observado. Assim, a LDA pode ser definida como:

> A alocação latente de Dirichlet (LDA) é um modelo probabilístico generativo de um *corpus*. A ideia básica é que os documentos sejam representados como misturas aleatórias sobre tópicos latentes, onde cada tópico é caracterizado por uma distribuição sobre palavras. (BLEI; NG; JORDAN, 2003, p. 996, tradução nossa).[21]

[21] Latent Dirichlet allocation (LDA) is a generative probabilistic model of a *corpus*. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words (BLEI; NG; JORDAN, 2003, p. 996).

Esse modelo probabilístico propõe modelar um conjunto de tópicos com base num contexto textual, dessa forma, as probabilidades de surgirem um tópico fornecem uma representação semântica do documento.

Em seguida, com o propósito de obter uma construção esclarecedora do que foi obtido no quadro, descreveram-se as principais características observadas em cada artigo analisado.

No primeiro artigo, foi proposta a implantação do método ***Hot Topics Rising Star Rank*** (HTRS-Rank) para encontrar acadêmicos iniciantes que contribuem com tópicos importantes no início de suas carreiras e os classifica com base na presença de tópicos importantes em suas publicações. Enquanto, na segunda pesquisa, do quadro, investigaram-se publicações científicas sobre Células Solares Fotovoltaicas Orgânicas e Inorgânica, utilizando-se de métodos bibliométricos para desvendar as possíveis tendências de pesquisa, e o crescimento de determinados temas.

Na terceira pesquisa, foi proposto um método de modelagem de tópicos para obter uma melhor representação de termos que forneça um direcionamento para pesquisas futuras em análise de tendências. Já a quarta pesquisa realizou uma revisão bibliométrica com o intuito de fornecer uma compreensão completa das tendências e tópicos relacionados ao tema "análise de sentimento", contribuindo no monitoramento eficiente dos assuntos que envolvem o tema.

A seguir, o quinto artigo analisado objetivou fornecer aos leitores uma visão abrangente do trabalho acadêmico sobre o tema "depressão de cuidadores" utilizando a bibliometria e a MT. As principais considerações revelaram que a prevenção ou controle da depressão entre cuidadores é um campo em crescimento com a maior prioridade para o envelhecimento da população. Assim foi verificada a contribuição deste estudo no campo do sofrimento psíquico dos cuidadores. Em seguida, no sexto artigo, foi realizado um estudo que obteve dezenove conjuntos temáticos, que foram encontrados e analisados, rotulados e sistematizados em quatro áreas principais: processos, tecnologias da informação, bibliotecas e documentações especializadas. Portanto, os pesquisadores identificaram as tendências globais sob o estado da arte analisado.

O sétimo artigo identificou diversas correlações temáticas que revelaram diversas distribuições de tópicos em países/regiões influentes e institutos de pesquisa. Essas descobertas ajudaram a compreender melhor a pesquisa científica sobre tecnologias utilizadas no cérebro humano, auxiliadas pela Inteligência Artificial. O oitavo trabalho utilizou-se de um estudo que avalia o uso de três diferentes abordagens reprodutíveis para identificar o surgimento de novidades tecnológicas em publicações científicas. Os resultados

sugerem que a contagem de termos produz resultados práticos para fins operacionais, enquanto o LDA oferece uma visão estratégica.

O nono artigo buscou fornecer orientações de pesquisas futuras com relação aos principais tópicos, temas e padrões de coautoria, realizou um mapeamento por meio de uma análise bibliométrica para visualizar o panorama da pesquisa. Além disso, identificou os temas subjacentes, as principais tendências e padrões e as futuras trajetórias de desenvolvimento de pesquisas foram mapeadas. Em seguida, o décimo artigo trata de uma pesquisa que obteve, como principais resultados, os índices de produtividade média dos pesquisadores analisados. Além disso, foram elaborados mapas de tópicos temáticos e gráficos de coautores, sugerindo as correlações das comunidades analisadas. A pesquisa concluiu ainda que o cruzamento de dados torna visíveis padrões subjacentes de hábitos de publicação, imperceptíveis nas estratégias metodológicas tradicionais.

No decimo primeiro artigo observado, foi possível verificar a utilização uma abordagem que identificou unidades semânticas por agrupamento de vetores de palavras. Além disso, foram comparados e discutidas as vantagens e desvantagens do método proposto em relação aos métodos existentes, descobrindo que, ao introduzir o significado semântico das palavras-chave, o método adotado no artigo suporta uma análise de conhecimento de domínio mais eficaz. Em seguida, a décima segunda pesquisa, trata de um estudo que buscou oferecer uma visão geral do estudo bibliométrico no domínio da Biblioteconomia e da CI, com o objetivo de dar uma perspectiva multidisciplinar dos limites tópicos e das principais áreas e tendências de pesquisa. Os resultados quantitativos revelaram a existência de 19 temas importantes, que podem ser agrupados em quatro grandes áreas: processos, tecnologia da informação, biblioteca e áreas específicas de aplicação da informação.

A décima terceira pesquisa buscou analisar um conjunto de dados de patentes relacionadas a veículos para descobrir indicadores temáticos. Ao usar métodos de aprendizado de máquina, a pesquisa se concentrou em diferentes métodos podem produzir padrões de semânticos a partir dos dados diretamente obtidos. Por fim, discutiu-se em detalhes as possibilidades de usar abordagens de aprendizado de máquina para a aplicação dessa metodologia. Em seguida, a décima quarta pesquisa analisada propôs uma metodologia na qual foi obtida uma série temporal entre as frases-chave, periódicos e autores melhor classificados de acordo com sua frequência. Também se examinaram os padrões nos principais periódicos, identificando simultaneamente a probabilidade do tópico em cada período, bem como os principais autores e frases-chave. Os resultados indicaram

que, nos últimos anos, tópicos diversificados tornaram-se mais prevalentes e tópicos convergentes tornaram-se mais claramente representados, no campo da bioinformática.

Os resultados, da décima quinta pesquisa, implicaram na criação de dois novos indicadores bibliométricos, incluindo proporção e tendência do tópico, para descrever as preocupações acadêmicas da temática investigada. Os resultados demostraram que a abordagem de modelagem de tópicos pode ser aplicada como estratégia metodológica em futuras avaliações de hidrelétricas e outros projetos de infraestrutura. Posteriormente, na décima sexta pesquisa foi realizada um estudo em que os autores adotaram uma abordagem para analisar a literatura sobre a doença de Alzheimer. Os resultados indicaram que o ano de 2013 foi o mais produtivo e o *Journal of Alzheimer's Disease* o periódico mais produtivo. Além disso, constataram-se 16 tópicos principais da literatura e verificou-se uma tendência de aumento perceptível em alguns tópicos.

Na décima sétima pesquisa, foi investigada uma abordagem para a sugestão e direcionamento em políticas de telecomunicações. Os resultados da análise de citações apresentaram que as publicações são geralmente citações recebidas, mas a maioria delas não recebeu citações altas, no campo da "política de telecomunicações". Outro ponto verificado revela que as publicações recentes não receberam grande números de citações e que a produtividade dos artigos, em termos de citações, aumentou nos últimos dez anos em comparação com as pesquisas anteriores a 2004. Logo após, na décima oitava pesquisa, os resultados da análise temporal mostraram que os recursos da área médica possuem propriedade mais atualizada do que os da área de informática e, portanto, de divulgação mais rápida ao público. Em segundo lugar, adotou-se uma medida de expoente e análise de conteúdo, para avaliar o método proposto. Com o método proposto, comprovaram-se como analisar campos temáticos diferentes de forma mais precisa e abrangente.

Finalmente, na última pesquisa investigada, os pesquisadores observaram a distribuição de autores e obras, a obsolescência e sua dispersão, e a distribuição da literatura por tema, ano e tipo de fonte. Dessa forma, eles constataram que tem havido um interesse constante por parte dos pesquisadores sobre a IA. Verificaram que os tópicos mais estudados foram as técnicas e métodos empregados e os aspectos gerais da IA. E, além disso, averiguaram que o nível de produtividade dos autores se ajusta à Lei bibliométrica de Lotka e, que a dispersão da literatura é baixa.

Na seção seguinte apresenta-se a ferramenta metodológica escolhida para realizar as análises bibliométricas e relações temáticas adotadas nesta investigação. O *software*

apresentado já está sendo utilizado na Literatura da CI brasileira e corresponde à uma proposta viável de visualização de agrupamento temático e de estatística textual.

2.1.4 Definição sobre o software de Estudos Métricos – Iramuteq

De acordo com Ferreira e Corrêa (2018) o Iramuteq[22] é um software gratuito, desenvolvido originalmente no idioma francês, pelo cientista Pierre Ratinaud, em 2009, como ferramenta de organização de dados. Ele ancora-se no ambiente estatístico do *software* R (www.r-project.org) e na linguagem Python. Como afirmou Lima (2017, p. 98):

> o princípio de funcionamento do software IRAMUTEQ é que ele estabelece uma interface com o software R e prepara a análise multidimensional dos textos e questionários. A sua operacionalidade consiste na preparação dos dados e em escrever os scripts que são depois lidos e analisados pelo software R. Os dados mostrados são o resultado da ligação destas duas aplicações. Imediatamente após a abertura de um "*corpus*" de dados, o IRAMUTEQ cria um dossiê na mesma pasta que o ficheiro foi aberto (**** *quest_01). É nesse espaço que se encontram os resultados da análise dos dados realizado por este software, ou seja, na pasta que você salvou os dados tratados.

Nesse caso, após a devida configuração do *corpus* a ser utilizado, o *software* permite a viabilização de diferentes tipos de análise de dados textuais, como lexicografia básica, lematização, cálculo de frequência de palavras, análises multivariadas, análise pós-fatorial e análise de similitude. O intuito de utilizar esse *software* é possibilitar uma vasta análise estatística que possa, igualmente, oferecer recursos visuais que facilitem o estudo.

O Iramuteq foi utilizado em diversos estudos, dentre os principais, destacam-se os seguintes:

a) A dissertação da pesquisadora Lima (2017), da área de Matemática, que realizou um estudo investigativo sobre a utilização das tecnologias digitais de informação e comunicação pelos professores de matemática da rede estadual de educação de Goiás. Em sua investigação, a pesquisadora utilizou o Iramuteq para construção de um gráfico de "Árvore de Similitude", a partir das respostas dos questionários, fornecidas pelos professores, e construiu uma nuvem de *tags* dos termos mais utilizados numa pergunta feita aos entrevistados. Nesse cenário, Lima (2017) obteve dados de conexidade entre as palavras que foram utilizadas nas repostas, e pode dissertar sobre as razões que levaram ao quadro obtido no gráfico (FERREIRA; CORRÊA, 2018, p. 4439).

[22] Disponível em: http://www.iramuteq.org/.

b) Também foi apreciada a tese de doutorado de Ferreira Júnior (2017), da área da CI, na qual ele conduziu um estudo de redes de relacionamento conceituais. O objetivo do trabalho foi buscar a caracterização de conceitos para a análise de sistemas de informação em ambientes digitais. Para tanto, ele se utilizou de técnicas e medidas próprias dos estudos de redes, a fim de identificar e caracterizar tais conceitos. Por meio do Iramuteq, ele conseguiu formatar categorias temáticas, com frequências simples e relativas, em que ele pode exportar os dados por meio de gráficos de rede e identificar características de similaridade, medidas de conexão e densidade do gráfico. Os resultados alcançados permitiram a identificação das relações de proximidade dos termos, explicando os vínculos de conectividade e tornando possível a análise das relações e atribuições de padrões de comportamento nas respostas obtidas (FERREIRA; CORRÊA, 2018, p. 4440).

Em vista disso, a procura por estabelecer parâmetros e indicadores que representem uma realidade é determinante para expressar os domínios da produção científica. Portanto, a adoção de instrumentos, como o Iramuteq, que possibilitem a utilização de alguns dos procedimentos de visualização temática propostos, confere uma vantagem competitiva para subsidiar as ações almejadas nesta pesquisa. Nesse mesmo caminho, a DC surge como área promissora que pode facilitar a busca por uma representação, ao se apoiar em diversos instrumentos de extração de informação. Na seção seguinte serão esclarecidos alguns pontos pertinentes da DC em Texto.

2.2 Descoberta de Conhecimento em Texto

A sociedade da Informação produz uma quantidade inimaginável de documentos e textos em diversos tipos de registros de conhecimento. Para tanto, ela se utiliza das TIC’s, com o objetivo de otimizar a velocidade de produção dos conteúdos e permitir o acesso instantâneo por meio das redes de comunicação.

A busca pela rapidez no processo de investigação do conhecimento oculto ocorre pela dificuldade em encontrar a informação pelo método tradicional. Devido a esse cenário, Marcondes, Costa e Martins (2016, p. 171) relatam que:

> As bases de dados bibliográficas, pela quantidade de informações e pelo seu crescimento, só podem ser eficientemente gerenciadas através de métodos computacionais. No entanto o acesso e reuso deste conhecimento é bastante problemático. Além do formato textual dos artigos digitais, estas bases de dados, em sua maioria, não são interoperáveis. Para serem acessadas utilizam sistemas de recuperação da Informação tradicionais com algoritmos de recuperação baseados na pouco expressiva Lógica Booleana da década de 1970. Por sua vez o processamento do conhecimento contido no texto de

> artigos com vistas ao reuso - identificação de lacunas, contradições ou concordâncias no conhecimento de determinada área, validação dos resultados de uma pesquisa - é extremamente trabalhoso, por demandar leitura e processamento por humanos.

Haja vista tais problemáticas, surge a preocupação em estabelecer formas de desvendar o conhecimento camuflado dentro dessas bases de dados, com o objetivo de facilitar o processo de acesso aos estudos armazenados. Se estabeleceu outrora, os sistemas de busca tradicionais não conseguem cobrir toda a cadeia de camadas de documentos presentes em muitas bases de dados, tornando-se, então, em muitos casos, ineficiente.

Nessa mesma lógica, com o intuito de facilitar o processo de acesso ao conteúdo e de promover o alcance de determinado tipo de conhecimento pelo pesquisador, são desenvolvidas pesquisas e tecnologias que permitem o acesso à grandes quantidades de informação, em alta velocidade. De acordo com Sérgio, Silva e Gonçalves (2016, p. 88), a DC em texto surge como uma versão da DC em Bases de Dados, e pretende, como objetivo principal, buscar manusear a informação que não esteja estruturada. Os autores afirmam que "este processo tem como objetivo desvendar padrões e tendências, classificando e comparando os mais variados documentos" (SÉRGIO; SILVA; GONÇALVES, 2016, p. 88).

Uma das razões que justificam a busca por tornar o acesso ao conhecimento mais eficiente é a velocidade com que estratégias governamentais e políticas públicas devem ser criadas e mantidas em prol do desenvolvimento da pesquisa científica no país. Agências de fomento como a Capes e o CNPq necessitam investigar o que vem sendo produzido no Brasil, para estabelecer as estratégias anuais de fomento à pesquisa.

Trucolo e Digiampietri (2014, p. 87) identificam bem essa questão, quando afirmam o seguinte:

> Estratégias e políticas públicas têm sido inseridas no país para melhorar a qualidade e aumentar a produtividade da pesquisa científica. Muitas vezes essas políticas são escolhidas de acordo com áreas de pesquisa já consolidadas e populares, nas quais se sabe que haverá retorno, ou ainda, identificadas como tendências mundiais.

Dessa maneira, não só a título de obtenção da informação num documento científico, mas como forma de identificar as tendências nas publicações, e assim ter a possibilidade de planejar como as diretrizes das políticas de fomento podem ser construídas, é o que justifica o desenvolvimento de pesquisas com esse perfil. É a partir desse papel estratégico que a DC surge como uma ferramenta importante nesse processo.

Em um estudo pioneiro, os autores Frawley, Piatetsky-Shapiro e Matheus (1992, p. 58), estabelecem que "*knowledge discovery is the nontrivial extraction of implicit, previously*

unknown, and potentially useful information from data". Percebe-se, então, que apesar de a DC explorar a extração de informação implícita, não significa que o material obtido será relevante, logo, o fator humano é crucial para decidir o que é ou não relevante. Mais adiante, os autores continuam a discussão e mencionam que a busca por padrões num conjunto de dados não atesta, com 100% de certeza, o grau de confiabilidade necessário para que aquele padrão se torne conhecimento; muitos padrões podem ser captados e não se tornar um estudo comprovado, "o conhecimento é útil quando pode ajudar a alcançar uma meta do sistema ou do usuário. Padrões completamente não relacionados aos objetivos atuais são de pouca utilidade e não constituem conhecimento dentro da situação especificada" (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992, p. 59, tradução nossa).

Portanto, de maneira concreta, o conhecimento alcançado pelos mecanismos de busca não garante sua utilidade se ele não for efetivamente aplicado pelo usuário. Na tentativa de compreender como a relação entre a DC e a MT estão próximas, faz-se perceptível, na maioria dos estudos, que os sistemas e técnicas descritas na DC estão, em verdade, baseados em métodos de aprendizagem de máquina, que são aprimorados para lidar com a busca de conhecimento em bancos de dados. Então, um algoritmo de dados é criado com o objetivo de aprender em um conjunto de arquivos de registros de um banco de dados, para finalmente retornar uma declaração, por exemplo, um conceito, representando os resultados do aprendizado como saída (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992, p. 61).

Entretanto, como asseguraram Feldman e Dagan (1995, p. 112), "a preocupação concentra-se na aplicação de técnicas de aprendizado de máquina e análise estatística para a descoberta automática de padrões nas bases de conhecimento"; isto posto, concebe-se mais uma evidência de que a aprendizagem de máquina fornece um conjunto de parâmetros necessários, isto é, que devem seguidos no transcurso da DC.

Marcondes *et al.* (2016) afirmaram, também, que o ato de processar o conhecimento inserido no texto de um artigo possui o objetivo de identificar lacunas, contradições ou aquiescências no conhecimento de uma área, servindo ainda como instrumento de validação das metodologias desenvolvidas, diminuindo a necessidade dispendiosa e trabalhosa que é demandada pela ampla leitura e processamento da linguagem por humanos. Os autores apontam que a tarefa base para o processamento automático de textos é a Extração de Informação (EI), que consiste em encontrar informações específicas no texto dos documentos, e, portanto, é necessário diferenciá-la de outros métodos ou processos:

> Diferente da Recuperação da Informação (RI) que tem por objetivo encontrar textos e documentos relevantes, de acordo com a consulta do

> usuário, a EI trata de solucionar o problema de achar informações dentro dos textos. Difere, também, da PLN (Processamento de Linguagem Natural) porque é mais específico, visando a extrair determinados tipos de informação (obter informação pré-especificada), geralmente direcionada para extrair características do domínio (termos, objetos, entidades, relações) no qual o texto está inserido. A EI é diferente, ainda, da Extração de Conhecimento (Descoberta de Conhecimento – Knowledge Discovery in Databases – KDD), porque não visa deduzir regras (MARCONDES; COSTA; MARTINS, 2016, p. 186).

Nesse ínterim, o avanço das pesquisas em MT torna-se importante para subsidiar a aquisição de conhecimento por especialistas nesses ambientes. Como afirmaram os seguintes autores numa definição pertinente quanto ao conceito de MT:

> A mineração de dados significa procurar padrões nos dados. Da mesma forma, a mineração de texto trata da procura de padrões no texto: é o processo de analisar o texto para extrair informações úteis para fins específicos (WITTEN; FRANK; HALL, 2011, p. 386, tradução nossa).

Nesse contexto, o grande volume de informações textuais sendo produzido a cada instante exige a adoção de ferramentas que possam subsidiar a busca desses padrões de maneira rápida e precisa. A técnica de MT é capaz de fornecer essas informações auxiliando no suporte à tomada de decisão e explorando arquivos de difícil acesso. Rezende (2005, p. 338) afirma que a MT é "um conjunto de técnicas e processos que descobrem conhecimento inovador nos textos".

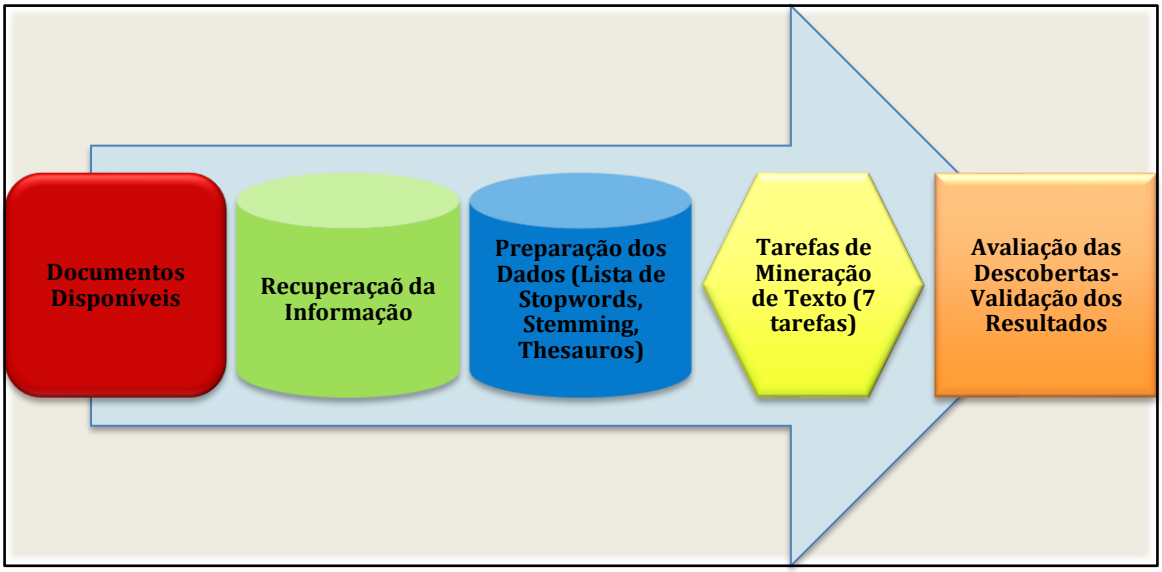
Outros autores, como Faro, Giordano e Spampinato (2012, p. 62, tradução nossa), declaram que "o principal objetivo da mineração de texto é descobrir o conhecimento oculto no texto em artigos publicados e apresentá-los aos usuários de forma coerente e concisa". Assim como a mineração de dados, a MT é tarefa específica da DC; ao passo em que Balaid *et al*. (2016) ressaltam que a MT pode ser considerada um processo de DC, adotando bancos de dados textuais ou a partir de recursos, utilizando de tecnologias. No mesmo período, os autores Nagarkar e Rumbhar (2015) afirmam que a MT está possuindo suas técnicas, utilizadas em muitas áreas do conhecimento, inclusive na Biblioteconomia e na CI, e que o seu principal objetivo é extrair arquivos em um conjunto de dados não estruturados.

Com o objetivo de extrair a informação, a ideia de minerar textos ou descobrir conhecimento neles pode também ser considerada um processo de descobrir padrões que sejam previamente desconhecidos, mas que tenham grande possibilidade de utilidade num grande conjunto textual. No entanto, o processo envolve uma série de interações do usuário com as ferramentas de MT, para explorar o repositório e encontrar esses padrões. Somente após essa interação e análise dos padrões, os especialistas podem determinar que

conhecimento extraído será útil para auxiliar na tomada de decisões (TSENG; LIN; LIN; 2007, p. 1219).

A Figura 4, a seguir, pode facilitar a compreensão dos processos envolvidos na técnica de MT:

Figura 4 – Etapas da Mineração de Textos



Fonte: (REZENDE, 2005, p. 339)

Observando a figura 4, identificam-se as cinco etapas do processo de aplicação da MT, percebe-se que desde a 1ª fase, da escolha e identificação dos documentos, até à avaliação dos resultados, percorre-se um complexo caminho que exige muita atenção e cuidado do pesquisador, principalmente nas etapas em azul e amarelo, onde competem a limpeza e a aplicação dos algoritmos escolhidos. Logo, mais especificamente na etapa correspondente ao hexágono em amarelo, identificam-se sete tarefas de MT, que podem ser adotadas no processo e serão descritas no quadro 7 a seguir:

Quadro 7- Tarefas de Mineração de Texto

TAREFA	FINALIDADE DA TAREFA
AGRUPAMENTO	Torna explícito o relacionamento entre documentos, agrupando documentos similares.
CATEGORIZAÇÃO	Identifica os tópicos-chave de um documento associando-o a categorias pré-definidas como áreas, domínios do conhecimento ou temas.
EXTRAÇÃO DE CARACTERÍSTICAS	Extrai padrões de termos com características em comum, também denominada extração de termos.
SUMARIZAÇÃO	Realiza o processo de redução textual no nível de sentenças, mantendo os significados-chave do texto
INDEXAÇÃO	Identifica os tópicos-chave de um documento por meio de padrões pré-existentes que permitem filtrar os termos que são palavras-chave de um documento.
REGRAS DE ASSOCIAÇÃO	Encontra relacionamentos ou padrões frequentes entre documentos, autores, citações, termos, categorias ou outros aspectos dos documentos.
REGRESSÃO	Gera um modelo preditivo da distribuição dos termos ou palavras-chave em um conjunto de documentos.

Fonte: (REZENDE, 2005)

Ao verificar o quadro acima, detectam-se os diversos tipos de tarefas de MT, segundo o autor supracitado. Percebe-se o grau de variabilidade de ação envolvida em cada uma delas e como isso pode influenciar a elaboração de cada estudo em específico. Nesta pesquisa, pretende-se adotar a "extração de características", a "indexação" e as "regras de associação", como tarefas da MT.

Seguindo esta lógica, apesar da busca pela automatização do processo de EI e DC, o componente humano ainda é muito importante, no que tange à identificação de informação relevante, como afirmou Srinavasan (2004, p. 403), que, em seu estudo, identificou que, apesar de desenvolver algoritmos específicos, com base no seu sistema de mineração de texto, para oferecer aos usuários especialistas em domínio um possível sistema de geração de hipóteses, a entrada do usuário é crucial para uma descoberta de conhecimento bem sucedida.

2.2.1 Estudos Sobre Mineração de Textos

Continuando a discussão iniciada na seção 2.2 e observando autores, na literatura científica, que se utilizaram da MT para obter êxito em seus resultados, inicialmente menciona-se a pesquisa de Trucolo e Digiampietri (2014); os autores buscaram identificar tendências das produções científicas de artigos de periódicos dos doutores da área de CI no Brasil. Eles utilizaram como fonte a Plataforma Lattes, e obtiveram 34.289 títulos de artigos publicados entre 1991 e 2012. O método consistiu em determinar palavras-chave mais

importantes a partir da extração automática de termos compostos inseridos nos títulos das publicações. Os principais resultados apontaram os temas mais trabalhados entre 1991 e 2012, e, por meio de um cálculo de regressão, foram projetadas tendências de frequência de ocorrência para os principais temas a serem trabalhados em 2013, 2015 e 2020.

Em seguida, apresenta-se o trabalho desenvolvido por Bezerra e Guimarães (2014); os autores buscaram demonstrar os termos mais empregados sobre trabalhos na área de Gestão do Conhecimento, ao longo de 10 anos. O estudo empregou 3.457 resumos de língua inglesa e 380 resumos em língua portuguesa no seu *corpus* de análise. Os principais resultados apontaram associações temáticas entre os termos presentes nos resumos dos artigos e identificaram os agrupamentos de termos mais frequentes e de maior ocorrência em relação ao total de documentos. Além disso, apresentaram termos mais associados aos aspectos pragmáticos e termos associados aos elementos mais abstratos da Gestão do Conhecimento.

Posteriormente, apresenta-se a pesquisa realizada por Araújo *et al.* (2016), intitulada Descoberta de Conhecimentos sobre Esquistossomose a partir de documentos científicos utilizando técnicas de Mineração de Textos. A investigação propôs a aplicação de técnicas de MT para a DC sobre esquistossomose a partir de um acervo científico do Instituto Oswaldo Cruz. Nesse estudo, os pesquisadores coletaram 179 resumos de artigos publicados entre os anos de 2005 e 2015, selecionados a partir do descritor “schisto”. Em vista disto, foram geradas listas de termos mais relevantes, documentos classificados por categorização e agrupamento sem suporte de especialista.

Adiante, evidencia-se uma pesquisa desenvolvida pelos autores Sérgio, Silva e Gonçalves (2016), que é publicada sob o título: Descoberta de Conhecimento a partir de informações não estruturadas por meio de técnicas de correlação e associação. Esse trabalho propôs a apresentação de um modelo para DC em textos com base nas técnicas de correlação e associação temporal entre termos de indexação.

Para isso, o *corpus* foi composto por artigos coletados na *Science Direct*, que foram utilizados para revelar as relações entre “Biotecnologia” e “Engenharia Genética”, e “Nanotecnologia” e “Medicina”. Foram coletados 551 artigos no ano de 2013, compreendendo 2 períodos para a busca, a saber: 1993 a 2002 e 1984 a 1993. Os resultados apontaram a evolução temporal dos relacionamentos entre os termos, possibilitando a identificação de padrões e tendências cuja proposta é conduzir à DC.

E, por último, destaca-se a nona pesquisa utilizada, produzida por Braga (2016). A autora propôs a criação de um modelo associado a uma representação de documentos por conceitos e aplicação de um método de agrupamento hierárquico de arquivos. Tal processo

metodológico baseou-se na frequência da ocorrência dos conceitos, com o objetivo de produzir uma taxonomia de conceitos. A pesquisadora utilizou o algoritmo, *a priori*, de Agrawal e Srikant (1994). O *corpus* analisado consistiu nos textos completos de 1.841 trabalhos científicos da biblioteca digital da Comissão Nacional de Energia Nuclear (CNEN). O principal resultado obtido foi uma árvore de taxonomia com a representação dos principais conceitos presentes nos trabalhos analisados.

Desse modo, dentre os procedimentos metodológicos que foram adotados nesse trabalho, uma parte muito importante corresponde às tarefas de MT descritas acima, na busca por compreender os níveis de produção e institucionalização das pesquisas sobre os temas escolhidos, e traçar um panorama da construção do conhecimento sobre esses temas em suas respectivas áreas. Na seção que segue, serão descritos os principais conceitos sobre a IA e como eles se relacionam com os outros temas discutidos nesta pesquisa.

2.3 Indexação Automática

Nesta seção serão apresentados os principais conceitos relativos à indexação, manual e automática. De maneira diacrônica, busca-se elencar os principais autores que marcaram as discussões iniciais em torno dos preceitos sobre indexação, para, posteriormente, abordar os principais conceitos sobre IA e apresentar suas características.

Uma discussão inicial sobre do que se trata a IA foi trazida por Maron em 1961. O pesquisador refletiu sobre a dificuldade existente em atribuir um conjunto de categorias e classificar corretamente o documento. O autor levantou um problema existente na atribuição de palavras para representar algo, e que a indexação de informações é algo complexo de se realizar.

> O termo "indexação automática" denota o problema de decidir mecanicamente a que categoria (assunto ou campo do conhecimento) pertence um determinado documento. Trata-se do problema de decidir automaticamente do que se trata um determinado documento. A situação é aquela em que existe uma coleção de documentos diferentes, cada um contendo informações sobre um ou vários assuntos. Também existe um conjunto de categorias, geralmente não exclusivas nem completamente independentes, mas (esperamos) exaustivas no sentido de que todo documento se "encaixará" em pelo menos uma das categorias fornecidas. O problema surge porque as categorias não são definidas extensionalmente. Ou seja, uma categoria não se determina enumerando todos e cada um dos documentos que constituem a sua composição, mas sim a situação inversa. Com base em alguma noção mais ou menos clara da categoria, devemos decidir se um documento arbitrário deve ou não ser atribuído a ela. Classificar corretamente um objeto ou evento é uma marca de inteligência; uma marca de inteligência ainda maior é poder modificar e criar categorias

> novas e mais fecundas, formar novos conceitos. (Talvez uma das características realmente dominantes de uma máquina inteligente seja a de criar novas categorias nas quais classificar suas "experiências".) Falando de maneira geral, parece que o problema de classificação tem duas partes. A primeira parte diz respeito à seleção de certos aspectos relevantes de um item como evidências. A segunda parte do problema diz respeito ao uso dessas evidências para prever a categoria adequada à qual o item em questão pertence. Mas antes de examinar essa maneira de encarar o problema, consideremos com mais detalhes o problema de classificar entidades linguísticas com base no que significam, em oposição ao problema de classificar as coisas em geral (MARON, 1961, p. 404, tradução nossa).[23]

Em consequência do que foi afirmado por Maron, já em 1961, a complexidade inerente ao processo de estabelecimento da busca e atribuição de decisão sob qual categoria atribuir, define melhor um determinado documento. Essa situação pode piorar num conjunto de documentos maior, contendo diferentes informações que precisam de um certo grau de exaustividade nas definições de determinadas categorias. Esse é um dos principais desafios, apontados pelo autor, da indexação automática.

Outro desafio inerente é a capacidade exaustiva em apontar ou indicar algo, como afirmou Stevens (1965), ele trata do substantivo *index*[24] como um significado geral servindo para apontar ou indicar algo. Assim o autor ainda afirma que um índice é um ponteiro que dirige o pesquisador para informações registradas. Neste sentido, a dificuldade em se estabelecer um "índice perfeito" é o maior desafio da indexação automática.

Em seguida, a título de buscar possíveis definições para o que seria indexação, assinala-se uma definição de indexação da UNISIST[25] (1975 apud Chaumier, 1988, p. 63),

[23] The term "automatic indexing" denotes the problem of deciding in a mechanical way to which category (subject or field of knowledge) a given document belongs. It concerns the problem of deciding automatically what a given document is "about". The situation is one in which there is a collection of different documents, each containing information on one or several subjects. Also there exists a set of categories, usually not exclusive nor completely independent, but (we hope) exhaustive in the sense that every document will"fit" into at least one of the given categories. The problem arises because the categories are not defined extensionally. That is to say, a category is not determined by enumerating each and every one of those documents which make up its membership, but, rather, the situation is reversed. Based on some more or less clear notion of the category, we must decide whether or not an arbitrary document should be assigned to it. To correctly classify an object or event is a mark of intelligence; a mark of even greater intelligence is to be able to modify and create new and more fruitful categories, to form new concepts. (Perhaps one of the really dominant characteristics of an intelligent machine will be that of creating new categories into which to sort its "experiences".) Loosely speaking, it appears that there are two parts to the problem of classifying. The first part concerns the selection of certain relevant aspects of an item as pieces of evidence. The second part of the problem concerns the use of these pieces of evidence to predict the proper category to which the item in question belongs. But before examining this way of looking at the problem, let us consider in more detail the problem of classifying linguistic entities on the basis of what they mean as opposed to the problem of classifying things in general (MARON, 1961, p. 404).

[24] Expressão do latim, também conhecida como agulha de bússola.

[25] *World Science Information System.*

nela a organização afirma que a indexação é a "operação que consiste em escrever e caracterizar um documento, com o auxílio da representação dos conceitos nela contidos". O autor Chaumier, ainda afirma que:

> Indexação é a parte mais importante da análise documentária. Consequentemente, é ela que condiciona o valor de um sistema documentário. Uma indexação inadequada ou uma indexação insuficiente representam 90% das causas essenciais para a aparição de "ruídos" ou de "silêncios" em uma pesquisa. Os 10 % restantes serão devidos a causas mecânicas tais como: erro de pontuação, de codificação, de transcrição etc. (CHAUMIER, 1988, p. 63).

Nesse âmbito, a indexação é um processo eficiente, capaz de oferecer qualidade no processo de busca de informação. Ainda nessa lógica, os mecanismos de recuperação da informação são beneficiados com uma indexação bem elaborada e precisa. Apesar disso, para se atingir um processo robusto de indexação é necessário o devido conhecimento do conteúdo do documento, pois, se os conceitos e nomenclaturas utilizados não forem dominados pelo indexador, dificilmente a tarefa se concretizará satisfatoriamente.

Para se obter o conhecimento mais profundo do documento, é necessário ter conhecimento sobre determinado tema ou assunto. E a forma mais comum em se conhecer um documento, seria lendo, contudo, o processo manual de leitura e atribuição de termos que representa o documento é, na sua origem, um processo demorado e custoso. Nessa linha, uma das maiores vantagens que se ganham ao se adotar a indexação automática, além do custo, é a unificação ou padronização dos índices gerados, como afirmaram Maarek, Berry e Kaiser (1991, p. 801, tradução nossa):

> A principal vantagem da indexação automática em relação à indexação manual, além das óbvias considerações de custo, é que ela permite um esquema unificado que garante que os índices sejam compatíveis entre si. A ideia é extrair atributos de uma fonte de informação existente; por exemplo: o código e a documentação em linguagem natural. Algum trabalho foi feito para a extração de informações funcionais primitivas do código; No entanto, a fonte mais rica de informações funcionais é a documentação em linguagem natural, supondo que haja alguma disponível.[26]

Isto posto, apesar da automatização ser importante para a construção de índices unificados, as fontes documentais, para a construção desses índices, continuam sendo os

[26] The major advantage of automatic indexing over manual indexing, besides the obvious cost considerations, is that it allows a unified scheme which ensures that indices will be compatible with each other. The idea is to extract attributes from an existing source of information; for example: the code and natural-language documentation. Some work has been done toward extraction of primitive functional information from the code; however, the richer source of functional information is the natural-language documentation, assuming that any is available (MAAREK; BERRY; KAISER, 1991, p. 801).

documentos elaborados por meio da linguagem natural. Essa linguagem, naturalmente, é mais rica por ter sido elaborada por humanos.

Continuando as discussões sobre os conceitos que envolvem a indexação, no Brasil, mais recentemente, Simões *et al*. (2017) afirmam que a indexação é uma operação que permite a descrição e a identificação do conteúdo de um documento por meio de termos previamente determinados. Todavia, para Borges e Lima (2015, p. 49) a indexação é vista como uma atividade:

> O processo de indexação corresponde à atividade de representar um documento através de uma descrição abreviada de seu conteúdo, com o intuito de sinalizar sua essência. Essa representação é feita a partir da análise de assunto do texto-fonte, que preferencialmente, deveria ser feita por especialistas da área, que tivessem um olhar atento para as metodologias e procedimentos provenientes da Ciência da Informação e da Biblioteconomia.

Em seguida, um outro conceito importante, mencionado por Lapa (2014, p. 60), é o de que 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 [...]"; neste caso, a indexação possui a capacidade de representar tematicamente o documento, pois ela o apresenta por meio de termos que descrevem seu conteúdo.

Outro conceito similar pode ser identificado na definição de Lima e Boccato (2009, p.136):

> A indexação é o processo de análise documentária que tem por finalidade identificar o assunto de que trata o documento e representá-lo através de descritores de uma linguagem documentária, de maneira a permitir a suja recuperação pelos usuários de um sistema de informação.

Portanto, observando as acepções, a indexação ofereceria um método aplicado e seguro para apresentar termos ou expressões que facilitam o processo de RI. De acordo com Pansani Junior e Ferneda (2016), a indexação é uma das etapas responsáveis por demonstrar os temas de um documento. Os autores afirmam que os termos possuem a capacidade de representar o conteúdo temático que servirá para recuperar um documento no momento da busca.

A partir deste ponto de vista, que tem por objetivo facilitar o processo de busca e permitir que o documento seja recuperado de maneira mais eficaz, a utilização do vocabulário controlado surge como uma oportunidade de proporcionar maior confiabilidade no processo.

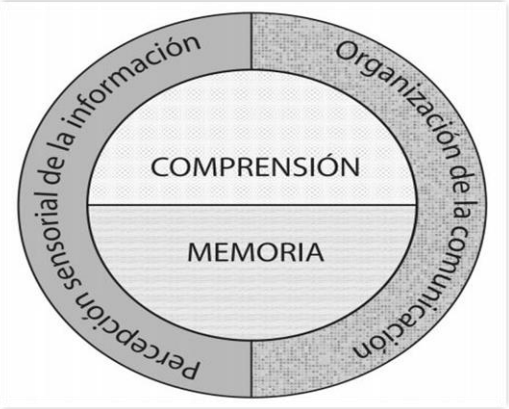
> Uma solução utilizada para promover uma maior uniformidade no processo de indexação é o uso de vocabulários controlados. Solução esta, que influencia positivamente nos processos de indexação e recuperação da informação, pois amplia o acesso às ideias do autor e serve como elo entre a

> informação criada e as necessidades dos indivíduos. A representação da informação realizada através da adoção de terminologias padronizadas resultou em avanços na qualidade da indexação. No entanto, o volume de documentos produzido pela sociedade ainda é um obstáculo para esta forma de indexação (PANSANI JUNIOR; FERNEDA, 2016, p. 4).

Neste sentido, percebe-se que o processo de indexar não é uma tarefa simples, entretanto, feito de forma adequada, com a utilização de vocabulário controlado, permite a diminuição dos erros e mitigação dos riscos que envolvem o processo de busca por documentos de cunho científico. Seguindo essa linha de pensamento, autores como Pulgarín e Gil Leiva (2004, p. 365, tradução nossa), afirmam que a "indexação é um processo intelectual que envolve leitura, compreensão, análise e representação"; nesse processo, uma das características mais evidentes é o componente subjetivo, que também pode ser chamada como "a subjetividade inerente à indexação". Isso se deve às diferentes interpretações, vindas de desiguais indexadores, ao examinar o mesmo documento (PULGARÍN; GIL-LEIVA, 2004).

Um modelo interessante que demonstra uma representação visual do processo intelectual adotado na indexação é empregado por Gil-Leiva (2008) no seu Manual de Indexação:

Figura 5 – Processo Cognitivo da Indexação



Fonte: (GIL-LEIVA, 2008)

Nesse modelo, percebe-se uma visão holística do autor quanto ao processo de indexação. Visualiza-se o indexador interagindo com a sua capacidade sensorial, identificando as estruturas textuais, por meio da compreensão, e raciocinando diante das memórias existentes. Posteriormente, ele realiza o processo de organizar aquilo que vai ser repassado. Nessa figura, constata-se que todo processo é contínuo e interligado.

Outra autora que trabalha conceitos pioneiros, já sobre automatização de resumos, é Jones (1974). Ela aponta a construção de resumos automáticos preparados pelo cientista Luhn em 1958, na Conferência Internacional sobre Informações Científicas, como o início dos

trabalhos que envolvem a indexação automática. Naquele momento, a autora supracitada produziu o índice *keyword in context* (KWIC) construído a partir de uma rotação automática das palavras significativas dos títulos. Ela afirma que a indexação diz respeito aos componentes linguísticos de uma recuperação de informação, por meio das linguagens natural e artificial.

Nesses termos, pode-se destacar um estudo de Névéol *et al*. (2008, p. 814), em que os autores afirmam ser a tarefa de atribuição de um número limitado de termos que denotam conceitos num documento, a principal definição para "indexação". Eles ratificam que a indexação é útil para fins de recuperação, possuindo, não obstante, uma forte relação semântica com o valor descritivo do documento; neste caso, um determinado conjunto de termos escolhidos possui a capacidade de detalhar um arquivo e servirá como uma sinopse do assunto discutido. Portanto, pode-se dizer que nesse processo "cada termo de indexação deve refletir um aspecto importante do documento e sua seleção constitui uma tarefa cognitiva difícil que implica uma compreensão completa do conteúdo do documento" (NÉVÉOL *et al.*, 2008, p. 814).

Na busca pelos conceitos sobre a indexação, identificam-se dois tipos existentes: a indexação manual e a indexação automática. Apesar de ambos utilizarem a compreensão textual como elemento motriz, existem diferenças específicas em cada uma delas, que podem ser descritas a seguir:

> **Tamanho das unidades documentais**: a indexação humana tende a se concentrar em unidades documentais maiores, como artigos de periódicos completos, capítulos completos em coleções ou mesmo monografias completas. Com a ampla disponibilidade de documentos de texto completo em bancos de dados de RI, a indexação automática agora recupera parágrafos individuais de textos, em vez de documentos completos. Mas é claro, os humanos poderiam, em princípio, analisar e indexar no nível de parágrafo da unidade documental, portanto, essa variável não está diretamente ligada ao método de indexação. **Extensão da matéria indexável:** vinculadas à disponibilidade de documentos de texto completo estão as diferenças na extensão da matéria indexável. A indexação automática agora é rotineiramente baseada no texto completo, enquanto grande parte da indexação humana pode ser limitada a um resumo ou outro resumo do texto completo. Essa diferença também está intimamente ligada à exaustividade, porque a exaustividade inferior da indexação humana típica pode ser acomodada com matéria indexável mais breve. Mas, novamente, os humanos podem analisar textos completos e indexar em níveis maiores de exaustividade, portanto, essas variáveis não são o mesmo que método de indexação. **Exaustividade:** a indexação automática tende a ser exaustiva, considerando a maioria, senão todas as palavras em matéria indexável como indicadores potenciais de conteúdo. Por outro lado, a indexação humana tende a ser seletiva, indexando apenas tópicos ou aspectos que parecem ser os mais importantes para resumir o conteúdo, significado ou propósito de uma mensagem. Mas, como acabamos de observar, os humanos podem

> analisar e indexar em níveis mais elevados de exaustividade, portanto, a exaustividade não está diretamente ligada ao método de indexação. **Especificidade:** a indexação automática tende a usar uma terminologia muito específica (e, portanto, um vocabulário muito amplo e variado), porque usa, ou pelo menos começa com a linguagem real do texto. A indexação humana tende a usar uma terminologia mais genérica (e um vocabulário muito menor em geral) na tentativa de resumir os tópicos e evitar a dispersão excessiva de tópicos intimamente relacionados. No entanto, os humanos podem usar vocabulários maiores, de modo que a alta especificidade não é necessariamente um atributo exclusivo da indexação automática (ANDERSON; PEREZ-CARBALLO, 2001, p. 234, tradução nossa, grifo do autor).[27]

No estudo em questão, os autores enumeram oito variáveis da recuperação da informação que devem ser consideradas ao se definirem as diferenças entre a indexação automática e a manual. As três variáveis escolhidas acima, refletem bem as principais diferenças encontradas naquela época em relação aos processos de indexação. O tamanho do documento é um dos fatores limitantes que pode influenciar a forma como a indexação será realizada. No início dos anos 2000, os humanos tinham a maior capacidade de indexar grandes documentos recuperados em bases de dados, contudo, pela própria limitação em razão da lentidão, não era possível para a indexação manual, e ainda não é, ser tão exaustiva. A exaustividade pode ser considerada uma característica principal da indexação automática. Contudo, os humanos são mais genéricos nos vocabulários utilizados no processo de indexação, utilizando uma quantidade menor de palavras, enquanto a automática, se utiliza de uma maior variedade dos termos adotados na indexação.

[27] Size of documentary units: Human indexing tend to focus on larger documentary units, such as complete periodical articles, complete chapters in collections, or even complete monographs. With the wide-spread availability of full-text documents within IR databases, automatic indexing now routinely retrieves individual paragraphs of texts, rather than complete documents. But of course, humans could in principle analyze and index at the paragraph level of documentary unit, so this variable is not directly tied to indexing method. Extent of indexable matter: Tied to the availability of full-text documents are differences in the extent of indexable matter. Automatic indexing is now routinely based on the complete text, whereas much human indexing may be limited to an abstract or other summarization of the complete text. This difference is tied closely also to exhaustivity, because the lower exhaustivity of typical human indexing can be accommodated with briefer indexable matter. But again, humans could analyze complete texts and could index at greater levels of exhaustivity, so these variables are not the same as indexing method. Exhaustivity: Automatic indexing tends to be exhaustive, considering most, if not all words in indexable matter as potential indicators of content. On the other hand, human indexing tends to be selective, indexing only topics or aspects that appear to be of most importance for summarizing the content, meaning, or purpose of a message. But, as just noted, humans could analyze and index at higher levels of exhaustivity, so exhaustivity is not directly tied to indexing method. Specificity: Automatic indexing tends to use very specificity terminology (and therefore a very large and varied vocabulary), because it uses, or at least begins with, the actual language of the text. Human indexing tends to use more generic terminology (and a much smaller vocabulary over all) in an attempt to summarize topics and to avoid too much scatter of closely related topics. However, humans could use larger vocabularies, so that high specificity is not necessarily an attribute unique to automatic indexing (ANDERSON; PEREZ-CARBALLO, 2001, p. 234).

Na Ciência Brasileira, um autor respeitado que discute de forma interessante os conceitos sobre a indexação automática é Robredo (1982). De acordo com ele, a indexação automática estabelece uma comparação de cada palavra do texto relacionando às palavras vazias de significado, que foram anteriormente escolhidas, sendo conduzido, assim, por eliminação, a considerar o conjunto de palavras que podem ser significativas no texto.

Já para pesquisadora brasileira, Vieira (1988, p. 48), "a indexação automática é uma operação que identifica, através de programas de computador, palavras ou expressões significativas dos documentos, para descrever de forma condensada o seu conteúdo". Destarte, Vieira (1988) ratifica o que foi proposto por Robredo (1982) ao identificar o programa de computador como definidor do processo de indexação. Ou seja, a criação de uma linguagem intermediária possibilita o reconhecimento do usuário, ou da máquina, por um texto que provavelmente não seria disponível para o usuário comum.

Entre os autores brasileiros, destaca-se, ainda, a definição de Corrêa e Lapa (2013, p. 258), quando estes afirmam que a indexação automática se apresenta como:

> [...] Um conjunto de operações realizadas pelo computador, de natureza estatística, linguística, ou de programação, destinado a selecionar termos como elementos descritivos de um documento pelo processamento automático de seu conteúdo.

Os autores citados constroem uma relação mais tecnicista no processo de atribuição de conceitos sobre a IA. Ao afirmarem como sendo um conjunto de operações informáticas ligadas à estatística, linguística e programação, percebe-se que o produto resultante deste processo está muito relacionado ao desempenho do sistema ou máquina utilizados.

Maron (1961), que é reconhecido por seu trabalho como pioneiro na discussão sobre IA, diz, em seu estudo intitulado: Indexação automática: uma investigação experimental, que busca relacionar as dificuldades da automatização na indexação porque a maioria das técnicas então existentes apontavam para características de frequência de termos num determinado documento, como sendo um fator decisivo no processo de escolha de um termo representativo. Além disso, ele aponta que a quantidade de palavras-chave existentes num documento está diretamente relacionada com o aumento da probabilidade de atribuições corretas de termos pela IA, ou seja, se um documento contém 10 palavras-chave, é bem provável que a IA seja mais eficiente em relação a um que possua apenas cinco.

Observando a partir de um prisma divergente, percebe-se o que é dito por Gil-Leiva (2017, p. 140), quando ele menciona a literatura se referindo à indexação automática de forma "variada". Em seu trabalho, são identificadas terminologias como: indexação assistida automatizada; indexação automatizada; indexação automatizada suportada; suporte

automático à indexação; indexação auxiliada por computador; indexação assistida por computador, dentre outros. Ainda assim, o termo que é mais encontrado na literatura é "indexação automática".

O mesmo autor ainda afirma que:

> A definição de indexação automática pode derivar de três perspectivas (Gil Leiva, 2008, p. 320): a) programas de computador que auxiliam no processo de armazenamento de termos de indexação, uma vez obtidos intelectualmente (indexação auxiliada por computador durante o armazenamento); b) sistemas que analisam documentos automaticamente, mas os termos de indexação propostos são validados e publicados - se necessário - por um profissional (indexação semi-automática); e, c) programas sem outros programas de validação, ou seja, os termos propostos são armazenados diretamente como descritores desse documento (GIL-LEIVA, 2017, p. 140, tradução nossa).[28]

Assim sendo, o referido autor contempla a IA com três definições que se conectam e podem fornecer uma compreensão facilitada do que seria esse processo.

Um dos textos mais citados na literatura internacional sobre IA é o artigo de Salton, Wong e Yang (1975). Nele, os autores elaboram uma relação matemática entre a densidade dos documentos pela sua posição vetorial e como isso se relaciona com a RI.

> Dois documentos, com termos de índices semelhantes, são representados por pontos muito próximos no espaço em geral, então a distância entre dois pontos do documento no espaço é inversamente proporcional com a semelhança dos vetores correspondentes. Como a configuração do espaço do documento é uma função da maneira pela qual os termos e pesos dos termos são atribuídos aos vários documentos de uma coleção, pode-se perguntar se um espaço ideal para documentos existe, ou seja, uma configuração que produz um ótimo desempenho de recuperação (SALTON; WONG; YANG; 1975, p. 613, tradução nossa).[29]

Os autores buscaram demonstrar a possibilidade matemática de tornar-se um grupo de documentos mais fáceis de recuperar. Nessa circunstância, eles comprovam que é possível alterar, ainda que artificialmente, as configurações do espaço do documento, a fim de gerar mudanças na RI. Um dos critérios apresentados pelos cientistas assegura que, numa coleção

[28] The definition of automatic indexing can derive from three perspectives (Gil-Leiva, 2008, 320): a) computer programs that assist in the process of storing indexing terms, once obtained intellectually (computer aided indexing during storage); b) systems that analyze documents automatically, but the indexing terms proposed are validated and published—if necessary—by a professional (semi-automatic indexing); and, c) programs without any further validation programs, i.e., the proposed terms are stored directly as descriptors of that document (automatic indexing) (GIL-LEIVA, 2017, p. 140).

[29] Two documents, with terms of similar indexes, are represented by very close points in space in general, so the distance between two document points in space is inversely proportional to the similarity of the corresponding vectors. As the document space configuration is a function of the way in which terms and term weights are found for the various documents in a collection, it may be asked whether an ideal space for documents, that is, a configuration that produces optimal performance recovery (SALTON; WONG; YANG; 1975, p. 613).

de pequenos documentos agrupados, e com ampla separação dos grupos individuais, eles podem apresentar melhor desempenho. Então, para obtê-lo, seria interessante aumentar a semelhança de documentos em diferentes agrupamentos.

Entretanto, a razão pela qual se deve adotar uma postura cautelosa ao se abraçar a indexação automática em detrimento à manual é pela possibilidade de imprecisão do processo. Jones (1974, p. 397) já afirmava, naquela época, dos obstáculos inerentes na utilização da indexação automática e como a avaliação equivocada pode colocar dúvidas na sua eficiência.

> A indexação automática pode ser avaliada de duas maneiras. Pode ser avaliado externamente, por comparação com a indexação manual. Ou uma técnica de indexação automática pode ser avaliada internamente, em comparação com outra. Podemos também distinguir avaliação macro e micro. Na avaliação macro, o desempenho de sistemas inteiros é comparado e, na avaliação micro, apenas os diferentes valores de uma determinada variável. Assim, na comparação da indexação automática e manual, a avaliação macro estaria envolvida se a indexação manual usando um dicionário de sinônimos fosse comparada com a extração e classificação automática de palavras-chave, e a micro avaliação se as técnicas manuais e automáticas para extração de palavras-chave fossem comparadas. Em comparações internas, a avaliação macro é ilustrada por comparações entre palavras-chave automáticas e classificação automática de documentos, e diferentes abordagens para analisar textos de entrada. No nível mais detalhado, é quase impossível limitar as comparações às mudanças no valor de um único parâmetro, uma vez que os processos que afetam uma variável provavelmente afetarão outras, mas a forma geral de avaliação é geralmente bastante clara. Deve-se admitir que a avaliação de recuperação costuma ser inadequada. Comparações entre os resultados obtidos por diferentes projetos também são dificultadas pela variedade de procedimentos de medição adotados. A recordação e a precisão são, no entanto, amplamente utilizadas e qualquer tentativa de comparar os achados nesta área deve referir-se a estas medidas, quaisquer que sejam as suas limitações teóricas. A avaliação da indexação automática é, portanto, principalmente em termos de desempenho de recuperação. Isso deve ser enfatizado porque nas primeiras tentativas as técnicas automáticas para, digamos, construir um tesauro, eram como alguém se produzissem grupos de palavras que teriam sido produzidos por um ser humano. Mas não há nenhuma boa razão para supor, se as técnicas automáticas e manuais geram resultados diferentes, que o desempenho do produto automático será inferior (JONES, 1974, p. 397, tradução nossa).[30]

[30] [30] Automatic indexing can be evaluated in two ways. It may be evaluated externally, by comparison with manual indexing. Or one automatic indexing technique may be evaluated internally, by comparison with another. We can also distinguish macro and micro evaluation. In macro evaluation the performance of whole systems is compared, and in micro evaluation only different values of a particular variable. Thus in comparing automatic and manual indexing, macro evaluation would be involved if manual indexing using a thesaurus was compared with automatic keyword extraction and classification, and micro evaluation if manual and automatic techniques for extracting keywords were compared. In internal comparisons macro evaluation is illustrated by comparisons between automatic keyword and automatic document classification, and micro evaluation by different approaches to parsing input texts. At the most detailed level it is almost impossible to confine comparisons to changes in the value of a single parameter, since processes affecting one variable are likely to affect others, but the general form of evaluation is usually clear enough. It must be admitted that retrieval evaluation is often

Percorrendo essa linha de pensamento, Salton e Yang (1973), argumentam que existem duas noções fundamentais que devem ser compreendidas ao se buscar realizar a IA. Os autores refletem sobre as diferenças e dificuldades existentes entre a exaustividade da indexação e a especificidade do termo. Pois, quanto mais se busca uma exaustividade provavelmente haverá dificuldade na descoberta de termos específicos, como colocam:

> A exaustividade da indexação se refere à precisão e profundidade com que as várias áreas de tópico pertinentes a um determinado documento são refletidas no conjunto de termos de índice atribuídos ao documento, enquanto a especificidade do termo é uma função da exatidão com que um termo caracteriza um determinado assunto. Em geral, aumentar a exaustividade implica em um melhor desempenho de recall, enquanto aumentar a especificidade do termo significa melhor precisão. Em particular, quanto mais exaustiva for a indexação, ou seja, quanto mais abrangente for a cobertura das várias áreas temáticas, mais provável é que os itens relevantes sejam realmente recuperados em resposta às consultas dos utilizadores, obtendo-se, assim, uma elevada recordação; da mesma forma, quanto maior a especificidade do termo, ou seja, quanto mais precisa a definição de cada termo, menos provável é que itens estranhos não relevantes também sejam recuperados, obtendo-se assim alta precisão. Em um determinado contexto de usuário e coleção, deve-se procurar um nível ótimo de especificidade no vocabulário e um nível ótimo de exaustividade na indexação para cobrir o desempenho de recall e / ou precisão desejado pela população de usuários (SALTON; YANG, 1973, 351). [31]

Como pode ser observado, os pesquisadores colocam que o incremento da exaustividade implica numa melhor recuperação, assim como o aumento na especificidade dos termos provoca uma maior precisão. Então, quanto mais exaustiva for uma indexação, mais temas ela conseguirá cobrir, porém, quanto mais específica for uma indexação, mais

---

inadequate. Comparisons between the results obtained by different projects are also made more difficult by the variety of measurement procedures adopted. Recall and precision are nevertheless very widely used, and any attempt to compare findings in this area must refer to these measures, whatever their theoretical limitations. The evaluation of automatic indexing is thus primarily in terms of retrieval performance. This must be emphasized because in early experiments automatic techniques for, say, constructing a thesaurus, were regarded as satisfactory if they produced groups of words which would have been produced by a human being. But there is no good reason to suppose, if automatic and manual techniques generate different outputs, that the performance of the automatic product will be inferior (JONES, 1974, p. 397).

[31] Indexing exhaustivity refers to the accuracy and depth with which the various topic areas germane to a given document are reflected in the set of index terms assigned to the document, whereas term specificity is a function of the exactness with which a term characterizes a given subject. In general, increasing exhaustivity implies a better recall performance, while increasing term specificity means better precision. In particular, the more exhaustive the indexing, that is, the more thorough the coverage of the various subject areas, the more likely it is that relevant items are actually retrieved in response to user queries, thus achieving high recall; similarly, the greater the term specificity, that is, the more precise the definition of each term, the less likely it is that extraneous non-relevant items are also retrieved, thus achieving high precision. In a given user and collection context, one must then look for an optimum level of specificity in the vocabulary, and an optimum level of exhaustivity in the indexing to cover the recall and/or precision performance desired by the user population (SALTON; YANG, 1973, p. 351).

termos relevantes ela vai recuperar, eliminando palavras que não serviram ao propósito da indexação.

Um exemplo que pode ser mencionado para elucidar essa questão é um estudo desenvolvido por De Winter, Zadpoor e Dodou (2014). Nele se identificou que os documentos indexados pelo Google Acadêmico, podem não ser, necessariamente, documentos científicos. Pois, diferentemente dos documentos indexados pela *WoS*, o google indexa relatórios, pré-impressões e teses, além de documentos com erros nos metadados, como pode ser observado:

> A indexação é feita automaticamente por analisadores que identificam dados bibliográficos nos documentos selecionados. Argumenta-se que, por causa de seu processo de inclusão automática, o Google é suscetível a erros nos metadados e na indexação de trabalhos não científicos (DE WINTER; ZADPOOR; DODOU, 2014, p. 1548, tradução nossa).[32]

Mais adiante, com o propósito de trazer outros tipos de pesquisas, mais recentes, que vem sendo desenvolvidas a nível internacional, buscou-se trazer alguns trabalhos que tratam da indexação automática ou extração automática de termos ou modelagem de tópicos. Neste sentido, é possível identificar algumas pesquisas que procuram extrair palavras automaticamente num conjunto de documentos. Como pode ser verificado a seguir:

a) A primeira pesquisa identificada, é a de Ni, Li e Chang (2020). Nesta pesquisa, os autores buscam elaborar um método supervisionado de classificação de palavras-chave baseado na tecnologia de extração automática de palavras-chave *TextRank*. Os autores buscaram otimizar o modelo com o algoritmo raiz, contribuindo com a modelagem das palavras-chave, além disso, cooperando com os temas atribuídos para a classificação de textos. A pesquisa propôs ainda a melhoria da precisão em comparação com os métodos convencionais de classificação e agrupamento e solucionou o problema sobre os métodos convencionais não terem os mecanismos de palavras-chave de auto-renovação e pesos de classificação de autoajuste.
b) Posteriormente, numa outra pesquisa, intitulada *Analysis of OWA (operator weighted average) operators for automatic keyphrase extraction in a semantic context*, dos autores Perez-Guadarramas *et al.* (2020), os pesquisadores apresentaram um método não supervisionado para extração de frases-chave, baseado no uso de padrões léxico-sintáticos para extração de informações de textos. Em consequência, eles utilizaram a abordagem da

[32] Indexing is done automatically by parsers that identify bibliographic data in the selected documents. It has been argued that because of its automatic inclusion process, Google is susceptible to errors in metadata and to indexing of no-scientific Works (DE WINTER; ZADPOOR; DODOU, 2014, p. 1548).

extração automática de frases-chave de textos com a perspectiva da análise semântica. A proposta de utilizar esse método é a de oferecer recursos linguísticos que permitam resultados com maior índice de acurácia e desempenho na recuperação de frases-chave.

c) Em seguida, relata-se a pesquisa dos autores paquistaneses Amin *et al.* (2020), eles desenvolveram um modelo de extração de frases-chave de textos da língua URDU[33]. A pesquisa propôs que modelo fosse capaz de identificar as frases-chave classificadas como tópicos. Assim, foram selecionadas as frase-chave com a pontuação mais alta como o tópico do documento. Os experimentos foram realizados em dois conjuntos de dados diferentes, e o desempenho do sistema proposto é comparado com as existentes técnicas de última geração. Os principais resultados demonstraram que o sistema proposto supera as técnicas existentes e possui a capacidade de produzir tópicos mais significativos.
d) O quarto trabalho identificado foi desenvolvido por Li, Hu e Chen (2020), os pesquisadores identificaram um problema na inspeção em projetos de construção civil. Tal inspeção gera grande trabalho ao se analisar uma grande quantidade de documentos, pois é preciso analisar grandes quantidades de dados textuais. Para solucionar este problema, os autores utilizaram uma nova abordagem de MT, baseada na extração de palavras-chave e modelagem de tópicos. Essa abordagem buscou identificar as principais preocupações e sua dinâmicas de problemas em 7250 documentos na área de construção civil. Nesse contexto, os resultados mostraram que o método proposto pode extrair com sucesso as principais preocupações ocultas em textos e identificar suas mudanças com o tempo, permitindo assim uma inspeção mais eficiente no local e uma melhor tomada de decisão.
e) Em seguida, relata-se a pesquisa desenvolvida por Perez-Guadarrama *et al.* (2018). Os autores buscaram desenvolver um método não supervisionado para extração de frases-chave de documentos de texto. Para tanto, eles combinaram o uso de padrões sintáticos-lexicais para identificar as frases candidatas, com um gráfico baseado em modelagem de tópicos, apoiado por um processo de análise semântica. A utilização de uma linguagem previamente definida permitiu o aumento na possibilidade de identificação das frases candidatas e a maior cobertura dos documentos textuais. Contudo, os autores afirmam que existe a demanda por melhoria nos resultados da extração de frases-chave em textos longos, e que, possivelmente, somente será solucionado, em futuras pesquisas.

[33] O urdu é uma língua indo-europeia da família indo-ariana que se formou sob influência persa, turca e árabe no sul da Ásia durante a época do sultanato de Deli e do Império Mongol (1200-1800). Isoladamente, urdu é o 5º idioma mais falado do mundo como idioma nativo, sendo o idioma nacional do Paquistão (WIKIPEDIA, 2021).

f) A sexta pesquisa a ser mencionada, é um trabalho elaborado por Barde e Bainwad (2017). Nesta pesquisa os autores enfatizam a importância de se estudar a modelagem de tópicos como sendo uma técnica poderosa capaz de analisar uma grande quantidade de documentos. A modelagem de tópicos pode ser utilizada como um instrumento para representar um conjunto de documentos por meio de palavras, categorizar textos, recomendar *tags*, extrair palavras-chave e filtrar informações. Os autores analisaram os modelos utilizados para realizar a modelagem, como o Modelo de Espaço Vetorial (VSM), o modelo de Indexação de Semântica Latente (LSI), o modelo de Indexação de semântica Latente Probabilística (PLSI) e o modelo de alocação de Dirichete Latente (LDA). Assim, os autores realizam uma análise de cada modelo levantando as vantagens e desvantagens de cada técnica de modelagem.

Por fim, para caracterizar a indexação automática é necessário mencionar seus dois tipos principais, que dependem da forma como os sistemas são elaborados. Lancaster (2004) promove uma descrição sobre os conceitos envolvidos na Indexação por Extração e por Atribuição. Segundo o autor, a Indexação por Extração pode ser denominada como extração automática na qual se extrai do texto os termos e logo após são realizados os cálculos de ponderação e selecionados os termos mais relevantes. Já na Indexação por atribuição, o autor descreve também como uma indexação por extração automática, neste caso, se associa um vocabulário controlado à um conjunto de termos ou expressões equivalentes que ocorrem com frequência nos documentos. Os termos serão selecionados após o cálculo de ponderação levando em consideração suas frequências no documento.

Com o objetivo de tornar mais claro o contexto sobre os dois tipos de indexação, apresenta-se o quadro 8 que esclarece os dois principais tipos existentes:

Quadro 8 – Tipos de Indexação Automática

Características da Indexação Automática por Extração	**Características da Indexação Automática por Atribuição**
Adota critérios de frequência, posição e contexto das palavras.	**Adota critérios de "perfil" de palavras ou expressões que costumam ocorrer frequentemente no texto.**
Utiliza o processamento do texto do documento para extrair termos com características estatísticas e probabilísticas.	**Consiste no processo de Representação temática do conteúdo do documento por meio de termos selecionados de um vocabulário controlado.**
É um processo menos complexo e não exige a utilização de vocabulário controlado ou controle terminológico.	**É um processo mais complexo de ser realizado por exigir um controle terminológico.**

Fonte: (BANDIM, 2017).

No quadro 8, é possível verificar as características específicas de cada tipo de IA; os 2 tipos apresentados, representam os mais comuns. Nesta pesquisa, foi adotada a IA por atribuição, por meio do *software* adotado nos procedimentos metodológicos, com o objetivo de fornecer maior eficiência na indexação.

Visando identificar os principais documentos que tratam sobre indexação automática na literatura internacional, optou-se por elaborar a subseção a seguir. Nela, serão descritas buscas realizadas na *WoS* sobre IA.

2.3.1 Observação do Estado da Arte

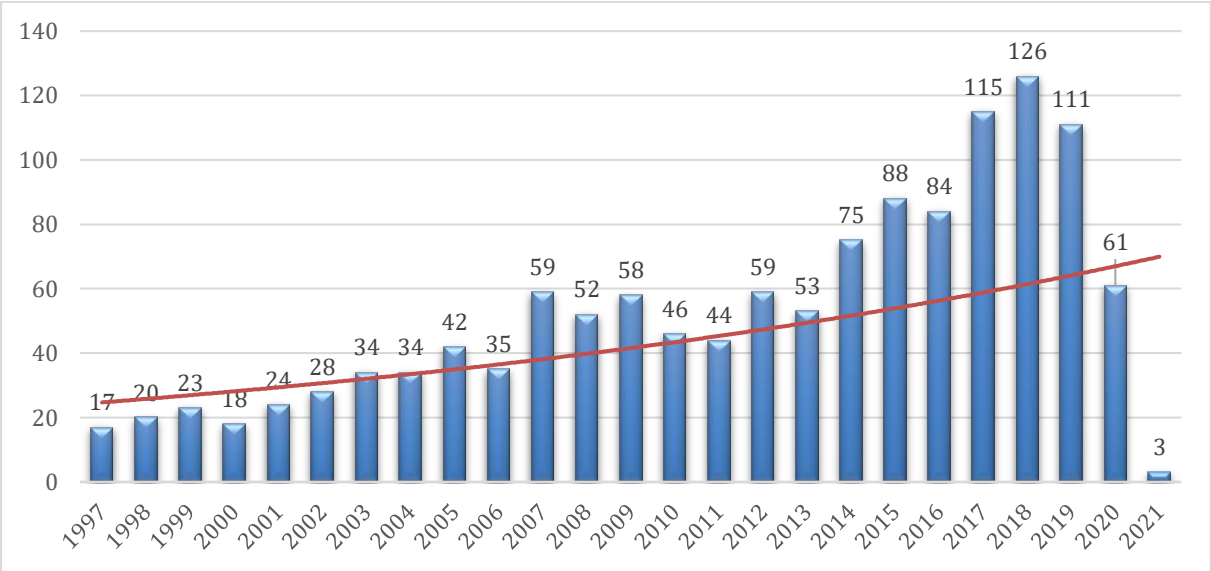
A partir do que se concebeu nesta subseção, será apresentado um conjunto de dados bibliométricos sobre o termo composto "indexação automática". Os resultados foram obtidos por meio buscas em três bases de dados internacionais.

a) Estudo na *Web of Knowledge*:

Na primeira delas, observam-se os dados obtidos na busca da *Web of Knowledge*. Para efeitos de observação e análise dos dados, foram realizadas apreciações numa planilha eletrônica. Para tanto, foi utilizada a seguinte expressão de busca na base: "*automatic indexing" OR "keyword extraction" OR "keyphrase extraction" OR "topic indexing"*.

Tal busca recuperou 1.482 documentos, no dia 09/12/2020, às 14:42h GMT -3. Os trabalhos encontrados contemplam publicações dos anos de 1945 até 2020. A seguir, observa-se um gráfico de tendência das produções ao longo de 25 anos, vide Gráfico 1.

Gráfico 1 – Tendência de produções de trabalhos ao longo de 25 anos



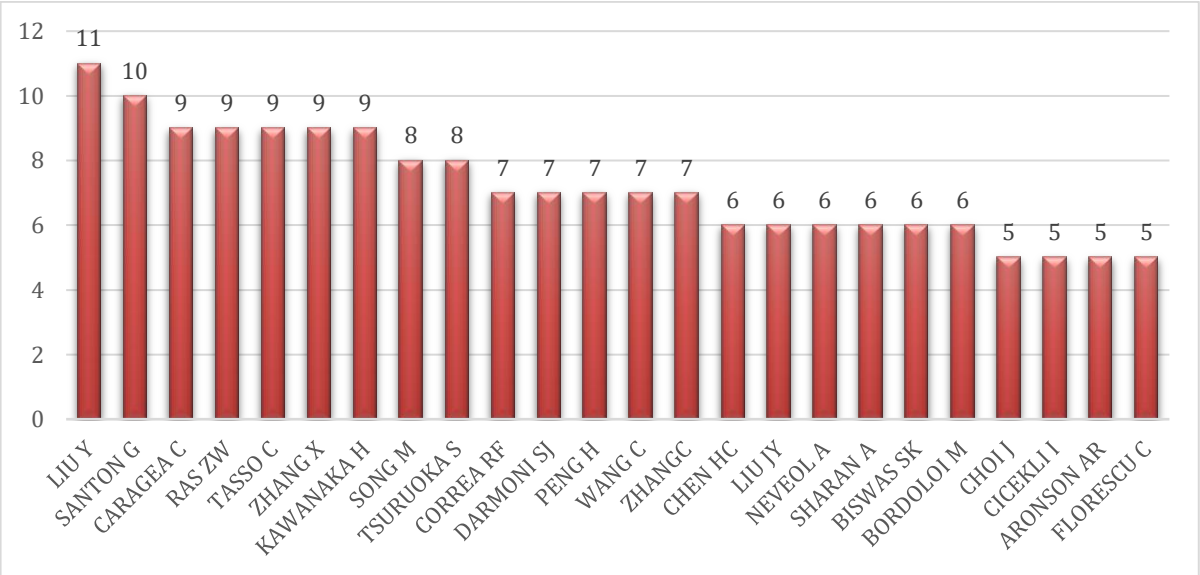
Fonte: *Web of Knowledge* (2020).

Neste gráfico, apresenta-se a quantidade de produções dos trabalhos indexados pela WoS desde o ano de 1997 até o ano de 2021. Ressalta-se que no ano de 2020 houve uma queda expressiva em relação aos anos anteriores, pois no período em que foi realizada a busca ainda restavam cerca de 21 dias para o término do ano; sendo assim, não era possível afirmar quantos artigos ainda teriam sido publicados sobre tais temáticas no referido ano. Porém, no gráfico, faz-se visível uma tendência de crescimento das produções ao longo do período, principalmente nos anos de 2017 e 2018, com 115 e 126 publicações cada. Percebe-se que nos primeiros 12 anos, isto é, no período compreendido entre os anos de 1997 e 2008, ocorreu a produção de 386 artigos, enquanto, entre 2009 e 2021, observa-se a produção de 923 artigos, o que corresponde a um crescimento de 239% no segundo recorte temporal contemplado.

Deste modo, concebe-se que as produções bibliográficas sobre IA, extração de palavras-chave, extração de frases-chave e indexação de tópico, vêm aumentando ao longo dos últimos anos, indicando, provavelmente, o interesse por pesquisas que envolvam o desenvolvimento de ferramentas e *softwares* de extração, assim como, pela sua potencial importância em contribuir com os processos de IA e EI.

Após o gráfico 1, é possível identificar os 25 autores com maior número de publicações desde 1963, que possuem trabalhos indexados na base, vide gráfico 2.

Gráfico 2 – Quantidade de trabalhos dos 25 autores mais produtivos



Fonte: *Web of Science* (2020).

Este gráfico contempla os autores que puderam contribuir de maneira quantitativa em relação ao índice de artigos publicados. Ressalta-se que foram utilizados dados dos 25 principais autores, pois o nível de dispersão é muito alto para se enumerar cada autor. Apesar do Autor LIU Y aparecer como o principal sobre o tema. Ao serem analisadas as 11 publicações e verificadas as autorias dos trabalhos, percebem-se variações no nome LIU Y abreviado, como: LIU YANG; LIU YUAN; LIU YU; LIU YE; LIU YING. Então, a abreviação Liu Y pode estar identificando trabalhos de autores diferentes para um mesmo autor.

Contudo, SALTON G aparece como o autor que possui maior quantidade de trabalhos sobre as temáticas, com 10 publicações. Ressalta-se que, dentre os 25 autores mais produtivos nas temáticas, apenas 1 é brasileiro, CORREA RF, com sete publicações sobre o tema.

Outro ponto importante a ser observado, é que dentre os autores que publicaram maior quantidade de trabalhos, alguns possuem os artigos mais citados na base, no gráfico 3 é possível analisar os índices de citação dos 20 artigos mais representativos. Ressalta-se que para esse gráfico, foram excluídos os dados de publicação dos seguintes periódicos ou congressos: Journal of Applied Crystallography; Metallurgical Transactions A-Physical Metallurgy and Materials Science; IEEE Transactions on Medical Imaging.

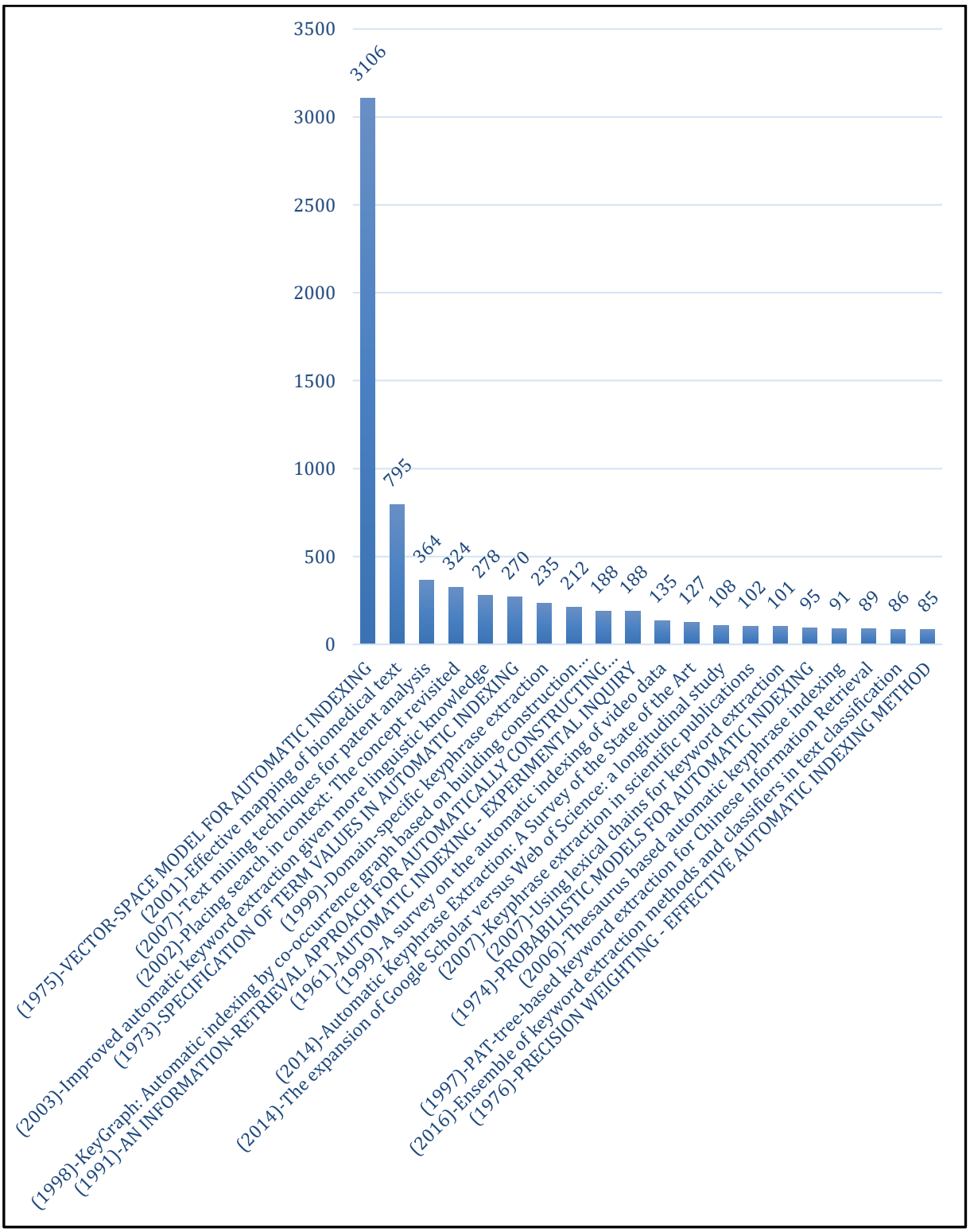
No gráfico 3, busca-se apresentar a quantidade de citações globais dos 20 artigos mais citados dentro da base. Percebe-se que o primeiro artigo apresenta 3.106 citações, dos autores (SALTON; WONG; YANG; 1975). Esse artigo destoa dos demais por apresentar um valor

muito elevado de citações, portanto, infere-se que o motivo principal seja o fato deles terem contribuído com o modelo de recuperação da informação mais utilizado atualmente. O número de citações desse artigo representa a soma das citações dos sete artigos seguintes, e, observando por esse ângulo, é possível refletir sobre a dimensão e o impacto gerados por esse *paper*.

Logo em seguida, nota-se os 19 artigos mais citados na base, publicados em diversas áreas do conhecimento, que tratam da temática sobre os tópicos buscados. Eles apresentam 3.873 citações no total. Uma das curiosidades sobre esse resultado é que: três artigos mais recentes que foram citados, são de 2014 e 2016; quatro são da década de 1970, inclusive o mais citado; um da década de 1960; cinco da década de 1990; e a maioria, sete, são dos anos 2000. Isso revela uma grande dispersão nos anos dos artigos que causaram maior impacto na base.

A partir das observações nos gráficos, pode-se perceber que o fato de possuir uma grande quantidade de artigos publicados, não influencia, necessariamente, na quantidade de citações. Outro ponto de vista a ser verificado é o tempo no qual o artigo foi publicado. Deduz-se que o artigo de Salton, Wong e Yang (1975) serviu de referência para os estudos na área da Indexação Automática, sendo base para muitos autores ao longo dos anos. Portanto, justifica-se o alto índice de citações direcionadas ao artigo mencionado. Contudo, de forma não menos importante, diversas obras contribuíram com a produção científica sobre o tema, como pode ser verificado no gráfico 3.

Gráfico 3 – Índice de Citações dos 20 artigos mais representativos



Fonte: *Web of Science* (2020).

Na tabela 1 é interessante notar quais são os autores que publicaram os artigos mais citados, representados no gráfico 3. Os autores destacados são os que possuem dois ou mais trabalhos indexados na lista dos mais citados. Infere-se que esses autores vêm se destacando

na produção nas temáticas que envolvem a extração de tópicos. Neste sentido, eles seriam os autores mais importantes em termos de impacto científico dentro da base *WoS*.

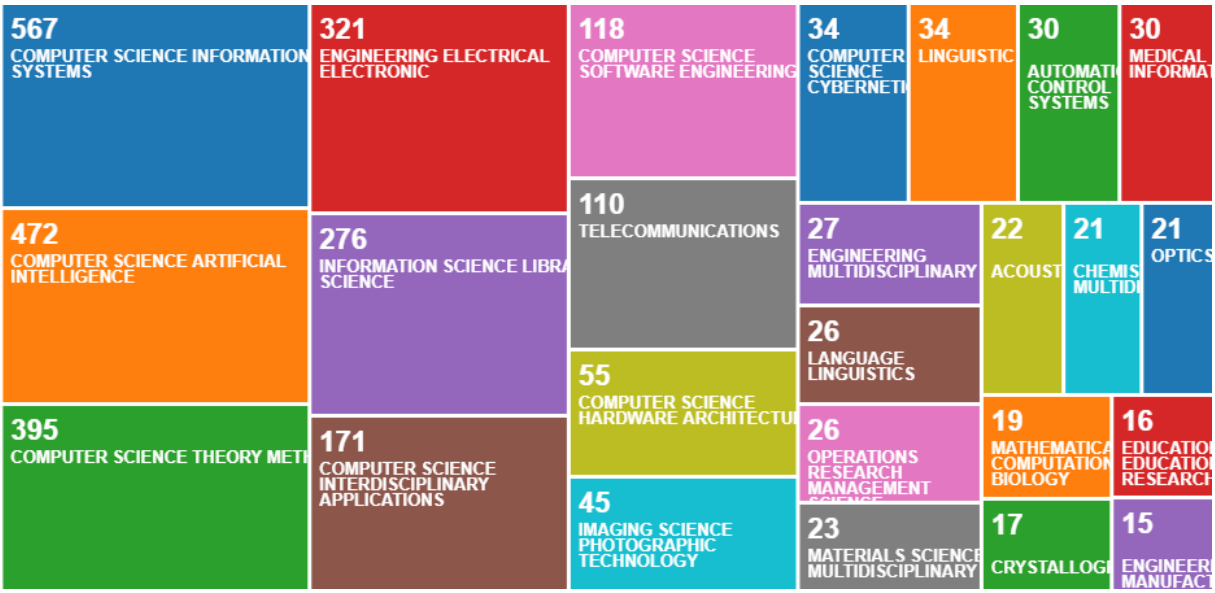
Tabela 1 - Autores dos 20 Artigos Mais Citados

AUTORES DOS 20 ARTIGOS MAIS CITADOS	**CITAÇÕES**
SALTON, G; WONG, A; YANG, CS	3106
ARONSON, AR	795
TSENG, YUEN-HSIEN; LIN, CHI-JEN; LIN, YU-I	364
FINKELSTEIN, L. *et al.*	324
HULTH, A	278
SALTON, G; YANG, CS	270
FRANK, E. *et al.*	235
OHSAWA, Y; BENSON, NE; YACHIDA, M	212
MAAREK, YS; BERRY, DM; KAISER, GE	188
MARON, ME	188
BRUNELLI, R; MICH, O; MODENA, CM	135
HASAN, KAZI SAIDUL; NG, VINCENT	127
DE WINTER, JOOST C. F.; ZADPOOR, AMIR A.; DODOU, DIMITRA	108
NGUYEN, THUY DUNG; KAN, MIN-YEN	102
ERCAN, GONENC; CICEKLI, ILYAS	101
BOOKSTEIN, A	95
MEDELYAN, OLENA; WITTEN, IAN H.	91
CHIEN, LF	89
ONAN, AYTUG; KORUKOGLU, SERDAR; BULUT, HASAN	86
YU, CT; SALTON, G	85

Fonte: *Web of Science* (2020).

O gráfico 4 apresenta uma lista das 25 áreas com maior quantidade de trabalhos indexados na *WoS*.

Gráfico 4 – Lista das 25 áreas do conhecimento com maior número de publicações



Fonte: *Web of Science* (2020).

Neste gráfico, identificam-se as 25 áreas que possuem maior quantidade de produções sobre as temáticas utilizadas nas buscas. O primeiro da lista, "*Computer Science Information systems*", apresenta-se como a área com maior quantidade de trabalhos publicados sobre o tema (567 trabalhos). Coincidentemente, o artigo que obteve maior índice de citações, como pode ser verificado no gráfico 3, também é dessa mesma área. Desse modo, sugere-se que a Ciência da Computação representa a área de conhecimento de maior impacto, no que diz respeito à IA e extração de tópicos.

Em seguida, percebe-se que as áreas que estão em segundo lugar (472 trabalhos) e terceiro (395 trabalhos) lugar também são relacionadas à Ciência da Computação, enfatizando a importância dessa área para o tema. Em 5º lugar, observa-se a CI como a área que mais publicou artigos com 276 trabalhos publicados. Dentre os artigos mais citados no gráfico 3, o artigo de Aronson (2001), com 795 citações, surge como um artigo indexado na área da CI. Neste sentido, percebe-se a Ciência da Computação como protagonista nas discussões e descobertas científicas da área, contudo, a CI, está intimamente relacionada às discussões sobre o tema e trabalha em paralelo com a Computação para promover as reflexões sobe a IA na WoS.

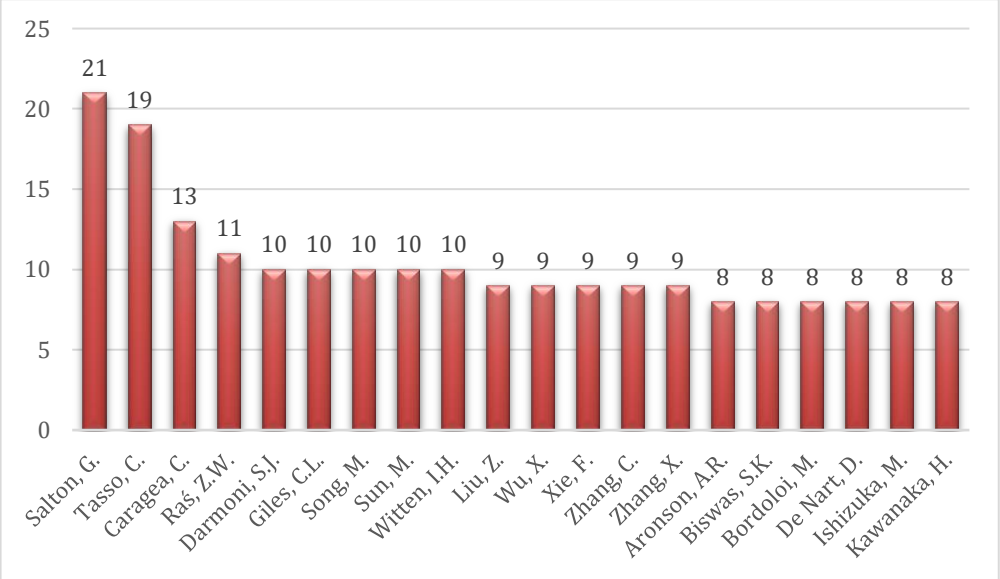
b) Estudo na *Scopus*:

Na segunda busca, observam-se os dados obtidos por meio da busca na base de dados *Scopus (Elsevier)*. Para efeitos de observação e análise dos dados, foram realizadas apreciações numa planilha eletrônica. Para tanto, foi utilizada a seguinte expressão de busca

na base: "*automatic indexing" OR "keyword extraction" OR "keyphrase extraction" OR "topic indexing".*

Tal busca recuperou 3.654 documentos, no dia 11/12/2020, às 10:20h GMT -3. Os trabalhos encontrados contemplam publicações dos anos de 1960 até 2021. A seguir, observa-se um gráfico de com os 20 autores com maior número de publicações desde 1960, que possuem trabalhos indexados na base, vide gráfico 4.

Gráfico 5 – Quantidade de trabalhos dos 20 autores mais produtivos



Fonte: *Web of Science* (2020).

Diante deste cenário, esta subseção buscou identificar quais os principais indicadores que expressam os domínios da produção científica sobre IA nas bases de dados mencionadas. Dessa forma, se fez possível identificar alguns dos principais autores ativos da área, produtores de artigos, que atuam desenvolvendo técnicas e *softwares*, com o propósito de tornar a extração de conceitos algo possível de ser obtido de forma eficiente e com alto nível de precisão.

Portanto, com o intuito de elucidar quais seriam os principais softwares de IA e Extração automática de termos, buscou-se realizar um estudo na *WoS* que proporcionasse um conjunto de informações sobre esses sistemas, e tais informações serão descritas na próxima subseção.

2.3.2 Identificação dos Sistemas de Indexação Automática ou de Extração de Palavras-chave

Aqui, pretendeu-se realizar um levantamento parcial das principais ferramentas identificadas no âmbito da IA no escopo internacional. Essas ferramentas são *softwares* que foram construídos com diversos propósitos como forma de extrair termos em diversos tipos de documentos e formatos. Para atingir os resultados obtidos neste quadro, foi realizada uma busca básica na *WoS* com os seguintes termos, separados pelo operador lógico OR, na opção tópicos: "*automatic indexing" OR "keyword extraction" OR "keyphrase extraction" OR "topic indexing"*. Essa busca recuperou 1.502 documentos, no dia 28/01/2021, às 09:53h. Para refinar a procura e selecionar os documentos, foram lidos os resumos dos 278 documentos indexados na categoria "*Information Science Library Science*".

Após a leitura, foi possível selecionar artigos que possuíam correlação temática com os termos utilizados na busca. Para encontrar esses artigos, foram baixados os metadados da *WoS* no formato de arquivo de texto sem formatação, e verificados os 278 registros de Títulos, Resumos e Palavras-chave. Como método de busca, foram utilizados termos como "*system*, *program* e *app*", com o objetivo de identificar algum desses sistemas nos metadados. Além desses artigos, com objetivo de mencionar sistemas importantes que não foram encontrados nessa busca, seja por limitações do filtro da área, ou por não ter sido possível encontrar nos metadados, buscou-se alimentar a tabela com informações de outros sistemas de IA, mencionados por Gil-Leiva (2017, p. 141). Os principais *softwares* identificados podem ser visualizados no Quadro 9.

Quadro 9 - Comparação entre os principais softwares de extração automática de termos

AUTORIA E ANO	OBJETIVOS DO SOFTWARE	NOME DO SOFTWARE E O *CORPUS*	CONSIDERAÇÕES E RESULTADOS SOBRE O TRABALHO
Kinglibiel (1973)	**O sistema indexa textos baseados num dicionário de palavras previamente armazenado. O Dicionário adotado possuía 21.500 unidades de palavras.**	**MAI (Machine-Aid Indexing)- 700.000 documentos técnicos do Centro de Documentação da Defesa (DDC)**	**Um dos maiores desafios levantados pelo autor é a dependência do sistema ao dicionário de palavras. Pois ele afirma que na medida em que o volume de texto aumenta é necessário atualizar o dicionário. Porém o sistema pode selecionar partes substantivas de títulos e resumos rapidamente, diminuindo a velocidade da indexação. O autor conclui que os sistemas conseguem indexar 1 milhão de palavras por hora. O sistema poderia**
Humphrey e Miller (1987)	**O sistema oferece a atribuição automática de palavras-chave de cabeçalho de assunto da área médica.**	**FrameKit- Utilizado para auxiliar indexadores da base de dados de saúde MEDLINE.**	**O sistema sugere ativamente relações associadas, que são os slots dos frames. Os indexadores inserem valores em resposta, o que por sua vez provoca o sistema para exibir quadros de instanciação adicionais.**
Knorz (1988)	**O sistema indexa textos baseados num dicionário de palavras previamente armazenado. O Dicionário adotado possuía 600.000 palavras-descritores.**	**AIR/PHYS – Base de dados da área de Física PHYS, contendo, ao final daquela época, mais de 1 milhão de documentos.**	**O processo foi feito com uma indexação semiautomática. Para cada 20.000 documentos indexados no sistema, 19% da indexação foi considerada boa pelos indexadores manuais. Contudo, os indexadores aceitaram 63% das indexações realizadas pelo sistema, considerando a necessidade de realizar algumas correções.**
Hersh e Greenes (1990)	**Utiliza um programa de computador experimental projetado para testar novas técnicas de recuperação da informação no campo da biomedicina. O**	**SAPHIRE (Ambiente de Recuperação de Informação Heurística Semântica e Probabilística) – Não foi utilizado *corpus* específico. Apenas adotaram-se alguns**	**Para este estudo, o SAPHIRE usa o algoritmo de lematização de Porter com o objetivo de indexar termos da área médica. O método foi aplicado simetricamente em todo o vocabulário para realizar o processo de indexação. O processo permite que o programa passe a lidar com diversos sufixos e plurais. Um algoritmo de retrocesso recursivo é então**

	software realiza descoberta de conceitos e tem a capacidade de processar textos livres para encontrar conceitos canônicos. O algoritmo é projetado para lidar com uma ampla variedade de sinônimos e convertê-los para a forma canônica. Isso permite que a linguagem natural possa ser usada para entrada e consulta, além de servir como base para uma nova abordagem para indexação automática baseada em uma combinação de métodos probabilísticos e linguísticos.	**experimentos usando termos da área de biomedicina.**	**realizado para encontrar os conceitos que englobam o maior número de palavras.**
Irving (1997)	**CAIT, o tutor de indexação assistida por computador, foi projetado para padronizar, agilizar e melhorar a qualidade de treinamento do indexador. Além disso, para suportar a necessidade crescente de todos os indexadores da Biblioteca Nacional Agrícola.**	**CAIT. A Biblioteca Nacional Agrícola Indexava cerca de 1.400 periódicos e 500 monografias sobre agricultura. A proposta do sistema é facilitar a indexação desses documentos, que levam, geralmente, de 12 a 18 meses com os indexadores manuais.**	**Oito indexadores avaliaram o "Módulo de Avaliação CAIT". Ele foi inicialmente desenvolvido para demonstrar os recursos do CAIT de uma forma abreviada. A maioria aprovou a aparência geral do CAIT e achou o desempenho satisfatório. Contudo, três avaliadores indicaram que se sentiram perdidos ou confusos sobre como utilizar o sistema. Seis dos avaliadores tiveram alguns problemas na compreensão das instruções dos exercícios. Contudo, a maioria dos indivíduos descobriu que os exercícios reforçaram os conceitos apresentados. No geral, os entrevistados sentiram que o**

			CAIT apresenta-se como uma ótima promessa.
Witten *et al.*, (1999)	**Utiliza aprendizagem de máquina para atribuir palavras-chave por meio de uma abordagem Bayesiana**	**KEA – 500 artigos de relatórios técnicos da ciência da computação**	**Kea pode, em média, identificar entre uma e duas das cinco palavras-chave possíveis escolhidas pelo autor. O software encontrou até 40% das palavras-chave do autor; Seu desempenho é suficiente para resumir, pesquisar e agrupar textos.**
Turney (2000)	**Usa um algoritmo de indução de árvore de decisão para classificar palavras como positivas ou negativas (C4.5) e usa um algoritmo genitor para ajustar parâmetros em um propósito específico (GenEx)**	**C4.5 e o GenEx – 290 textos de jornais e e-mails**	**O GenEx é melhor que o C4.5 pois o GenEx ao incorporar um conhecimento de domínio processual especializado, realiza um processo mais detalhado do que o C4.5. A precisão do software atingiu no máximo 30% das palavras escolhidas pelo componente humano. Porém, até 80% das palavras-chave escolhidas possuíam uma qualidade aceitável para leitores humanos.**
Barker e Cornacchia (2000)	**Utiliza um algoritmo (B&C) para extração de palavras-chave que examina um documento por meio de palavras nominais de base. Ele atribui pontuações a essas palavras, baseadas na frequência e no comprimento. E compara esse algoritmo com outro já utilizado por Turney (1999) – Extractor. Um grupo de pessoas (juízes) foi escolhido para avaliar as escolhas dos sistemas**	**B&C e Extractor**	**Os resultados demonstraram que as palavras-chave escolhidas pelo sistema B&C foram validadas em 40% das vezes pelos juízes, enquanto que o Extractor em 39%. As palavras-chave foram melhor atribuídas, quando escolhidas unicamente, pelo Extractor, contudo, o conjunto de palavras atribuídas pelo B&C foi melhor avaliado pelos membros que julgaram os resultados.**
Aronson (2001)	**Programa desenvolvido para**	**Metamap**	**Pesquisas mostraram que o MetaMap é uma ferramenta**

	mapear texto biomédico para o Metathesaurus ou, equivalentemente, para descobrir os conceitos de Metathesaurus mencionados no texto. O MetaMap utiliza uma abordagem intensiva em conhecimento, baseada em técnicas simbólicas de processamento de linguagem natural (PNL) e linguística computacional.		**eficaz para descobrir conceitos do Metathesaurus no texto. Mas existem duas áreas em que o desempenho do MetaMap requer melhorias: primeiro, a detecção de textos idiossincráticos como nomes químicos, acrônimos e abreviações, quantidades numéricas ou construções semelhantes; e segundo, resolução de ambiguidade.**
Montejo Ráez (2001)	**Consiste no desenvolvimento de um sistema de indexação automática por atribuição de tarefas. Realiza a seleção de palavras-chave, com a utilização de um vocabulário controlado, descrevendo e resumindo os conceitos considerados importantes no *corpus* analisado.**	**O sistema propõe a atribuição de palavras-chave a partir de artigos completos da língua inglesa sobre Física de Altas Energias. O servidor weblib.cern.ch cobria mais de 430.000 referências e 170.000 documentos no formato eletrônico, portanto, o uso de palavras-chave para classificação e busca é muito relevante.**	**Os testes iniciais foram realizados em uma coleção de 1.200 documentos. A partir disso os resultados apontaram um índice de precisão de 53,8% e um índice de revocação de 59,7%. Sua interface web permite a ampliação do uso da ferramenta e permitiu converter-se numa ferramenta de indexação valiosa para os pesquisadores do CERN.**
Chen (2005)	**Utiliza um algoritmo para extrair automaticamente palavras-chave de páginas da Web. O propósito é extrair**	**Kex algorithm – 381 páginas da web de www.msn.com**	**O Kex algorithm apresenta resultados melhores do que o Kea tendo um melhor desempenho na precisão. Os resultados apresentam que o software apresenta ser mais estável e recupera mais palavras-chave do**

	palavras relevantes sobre os principais tópicos tratados.		**que o Kea.**
Kolar, Vukmirovi´c, Ba˘si´c, and ˘Snajder (2005)	**Sistema de documento Auxiliado por Computador que aplica palavras-chave de vocabulário controlado do dicionário de sinônimos EUROVOC**	**CADIS**	**A principal contribuição deste artigo é a introdução da estrutura de dados interna CADIS especial que lida com a complexidade morfológica da língua croata. A estrutura de dados interna do CADIS garante estatísticas eficientes e análise de documentos de entrada e feedback visual rápido proporcionando a indexação de documentos mais rapidamente, com maior precisão e uniformidade.**
Kelleher e Luz (2005)	**Utiliza um recurso de Semantic Ratio (SR), aplicado ao Kea, que induz um aumento na extração de termos por meio da associação de estruturas semânticas.**	**KeaWeb -°**	**As aplicações iniciais demonstram que a estrutura de extração automática de frases-chave por meio do KeaWeb é 50% mais eficiente. Os resultados iniciais apresentam uma melhoria no sistema baseado em análise de hiperlink e podem trazer melhorias significativas na pesquisa da Web existente.**
Huang *et al.*, (2006)	**Propõe um software que extraí frase-chave automaticamente usando 2 novos recursos de pesos no Kea.**	**Kea-svm**	**O software se propõe a realizar extrações automáticas supervisionadas e não supervisionadas, tratando cada documento como um membro da rede semântica.** **Os experimentos demonstraram que o algoritmo de extração melhora em 50% a eficácia e 30% em eficiência em tarefas não supervisionadas. Na tarefa supervisionada a precisão ficou em até 80%.**
Medelyan e Witten (2006)	**Apresenta um algoritmo para uso de indexação de documentos, usando a indexação por extração de palavras-chave. Utiliza o vocabulário**	**Kea++**	**A principal constatação é que o Kea++ supera o Kea original em níveis de precisão 1,5 vezes maior. O Kea++ alcançou um percentual de 27% de consistência com os termos atribuídos pelos humanos que foram de 38%.**

	controlado, eliminando a ocorrência de palavras que não estejam no contexto, reduzindo a quantidade de treinamentos.		
Sinkkilä, Suominen e Hyvonen (2011)	**O objetivo foi testar a ferramenta de indexação automática avançada Maui em textos finlandeses, usando três algoritmos de *stemming* e de lematização. Os testes foram realizados com documentos e vocabulários de diferentes domínios.**	**Maui – 60 documentos extraídos aleatoriamente do repositório Sosiaaliportti, mantido pelo Instituto Nacional Finlandês de Saúde e Bem-Estar.**	**O teste de consistência entre indexadores constatou que a consistência entre os indexadores era de 33,7%, enquanto a consistência entre Maui e os indexadores humanos foi de 27,9%. Maui faz anotações em tópicos quase tão bem quanto os indexadores humanos. O desempenho de Maui chegou a ser ligeiramente superior em 1 dos 6 indexadores humanos.**
Chebil *et al.* (2012)	**Consiste num sistema chamado Catálogo e índice de sites médicos em língua francesa (CISMeF). Foi desenvolvido com a proposta de encontrar informações médicas úteis na Internet, destinadas aos profissionais de saúde.**	**O CISMef foi utilizado em 500 documentos da língua francesa para avaliar a indexação automática em documentos médicos. Foram indexados os campos dos títulos e subtítulos.**	**Os principais resultados apresentaram uma medida de precisão de (0,56) para títulos curtos e legendas, e (0,39) para títulos longos e subtítulos. O aumento do número de palavras por título e subtítulo gerou essa diminuição no índice de precisão. As medidas de revocação também foram baixas independente do comprimento dos títulos e subtítulos. Portanto, os resultados concluem que a função de indexação automática utilizada pelo CISMeF é mais eficiente em frases simples e curtas.**
Salisbury e Smith (2014)	**Consiste num sistema chamado de AgNIC (Rede de Informação Colaborativa para a Agricultura).**	**O método verifica automaticamente os campos do Tesauros da biblioteca Nacional de**	**Os principais resultados apontam que o método de indexação semiautomática não se restringe a identificar um vocabulário controlado. O método pode ser utilizado em qualquer dicionário**

	Utiliza como suporte a Biblioteca Nacional Agrícola (NAL) com o objetivo de contribuir na indexação e elaboração de resumos de textos acadêmicos da área.	**Agricultura (NAL). O processo é semi-automático e serve para que um indexador decida quais os termos são mais aplicáveis num conjunto de metadados a ser inserido n registro.**	**de sinônimos que indexe qualquer campo de metadado.**
Gil-Leiva (2017)	**Trata do sistema SISA (Sistema para a Indexação Automática para Artigos Científicos), que adota regras de heurísticas de localização e regras estatísticas, como frequência de termos, para obter a indexação automática ou semi-automática dos textos.**	**O artigo buscou analisar 450 artigos e um total de 2.077 descritores de 3 bases de dados científicas espanholas. Dessa maneira, o objetivo da pesquisa era verificar quais regras (heurística de localização ou regras estatísticas) forneceriam os melhores termos indexadores de uma coleção de documentos.**	**Os principais resultados apontaram que dos 2.077 descritores analisador, 792 (38,1%) apareceram no título ou no resumo ou em ambos. Em contrapartida, o restante (61,9%), estavam em nenhum dos dois campos. Então, os resultados indicam que a indexação por regras de localização funciona melhor do que a por regras estatísticas. Contudo, apesar dos resultados, as de regra de localização exigem a rotulação do documento, gerando um processo demorado e dispendioso.**

Fonte: Dados da pesquisa (2020).

Os dados apresentados no quadro 9 demonstram a evolução gradativa ao longo de 11 anos da implementação de sistemas de IA em conjuntos de *corpus* diversos. Inicialmente, obtiveram-se nove trabalhos que continham ferramentas de IA, utilizadas nos procedimentos metodológicos dos artigos. É possível identificar que o *software* Kea[34] apresenta-se como o precursor, servindo, inclusive, de parâmetro para a implementação de novos *softwares*. Seu código aberto facilitou o desenvolvimento por outros pesquisadores, acarretando, consequentemente, na melhoria de suas funções em índices de precisão. No geral, percebe-se

[34] Software desenvolvido na universidade de Waikato (Nova Zelândia). É um algoritmo de aprendizagem para treinamento e extração de frase-chave. Pode ser encontrado no Projeto Biblioteca Digital (http://www.nzdl.org/).

que os índices de revocação na extração automática não ultrapassam os 50%; contudo, quando supervisionada pelo ser humano, ela obtém até 80% de precisão, conforme Huang et al. (2006).

Entretanto, o *software* que se apresenta como um dos mais próximos ao desempenho do ser humano é o Maui. Esse software, de código aberto, foi utilizado por três pesquisadores finlandeses e comparados a indexadores humanos. O principal propósito desses estudiosos era identificar se o nível de consistência apresentado pelos indexadores humanos se aproximava do nível de consistência do Maui. O que ficou evidente é que, apesar de possuir uma consistência ligeiramente menor, o *software* chegou muito próximo dos resultados alcançados pelos humanos, portanto, possuindo um bom nível de consistência.

Além disso, o *software* foi desenvolvido por uma cientista da Nova Zelândia, Olena Medelyan, em sua tese de doutorado. A pesquisadora construiu experimentos de extração automática de frases-chave na Universidade de Waikato, aplicando os métodos do *software* Kea à indexação de frases-chave com vocabulário controlado. Consequentemente, ela conseguiu desenvolver o Maui, um indexador de tópicos multiuso que possui as funcionalidades do Kea, e permite o uso da Wikipédia como um vocabulário controlado. Com isso, o *software* melhora o desempenho das tarefas de indexação de tópicos em comparação com o Kea (MEDELYAN, 2009).

Por esse motivo, e por possuir o código aberto, o software Maui foi escolhido, nesta tese, como o sistema a ser desenvolvido e utilizado nos experimentos que farão parte da IA nesta pesquisa.

Os resultados obtidos no quadro 9, podem ser comparados com os resultados obtidos da pesquisa desenvolvida por Shaikh (2018), no quadro 10, “Keyword Detection Techniques: A Comprehensive Study”. Neste trabalho o autor faz uma revisão das principais técnicas existentes para a detecção e extração de palavras-chave. O quadro 10, a seguir, exemplifica bem quais são as principais técnicas identificadas pelo autor.

Quadro 10 - Técnicas de detecção de palavras-chave descritas no artigo "Keyword Detection Techniques: A Comprehensive Study"

ARTIGO	**TÉCNICAS UTILIZADAS**	**VANTAGENS**	**DESVANTAGENS**	**RESULTADOS E ANÁLISES**
H. Zhao, Q. Zeng. (2013). ***"Micro-blog keyword extraction method based on graph model and semantic space".***	a) Graph model b) Semantic space	a) Can detect the words which are wrongly segmented. b) Extracts keywords from a micro blog.	a) Not suitable for large texts. b) Some terms will not be distinguished.	Best performance obtained is 0.6972
H. Hromic, N. Prangnawarat, I. Hulpus, M. Karnstedt, C. Hayes. (2015). ***"Graph-based methods for clustering topics of interest in Twitter".***	a) OSLOM algorithm b) Page rank algorithm	a) Able to identify the topics of twitter event. b) Less expensive	Not able to identify the events based on graph clusters.	Best result obtained from structured based approach
L. Marujo, W. Ling, I. Trancoso, C. Dyer, A. W. Black, A. Gershman, D. M. de Matos, J. P. Neto, J. G. Carbonell. (2015). ***"Automatic keyword extraction on Twitter".***	a) Brown clustering b) Continuous word vector	a) Improved state of the art for keyword extraction. b) Automatically keyword extraction	Not suitable for Facebook text keyword extraction.	Accuracy for precision obtained is 72.05, recall 75.16.
] P. Torres-Tramon, H. Hromic, B. R. Heravi. (2015). ***"Topic detection in Twitter using topology data analysis".***	TOPOL	a) Suitable for noisy data. b) Reduces computation time and improves topic extraction result	Suffers from data fragmentation.	The result obtained is 0.5380 for recall,0.7500 for precision.
W. D. Abilhoa, L. N. De Castro. (2014). ***"A keyword extraction method from Twitter messages represented as graphs",.***	a) Tf-Idf b) KEA c) Proposed TKG	a) TKG proved to be robust and superior compared to other approaches b) TKG is simpler to use than KEA	The best configuration of TKG was not found	TKG results better compared to KEA and Tf-Idf

W. Chung, H. Chen, J. F. Nunamaker Jr. (2005). "*A visual framework for knowledge discovery on the web: An empirical study of business intelligence exploration".*	a) Statistics approach b) Machine learning approach	a) Search engine which can automatically extract important keywords b) System works well	Not suitable for business management domain	High recall rate
D. Isa, L. H. Lee, V. P. Kallimani, R. Rajkumar. (2008). *"Text document preprocessing with the bayes formula for classification using the support vector machine",*	Bayesian approach	a) Low cost, simple and efficient method. b) Handles raw data without text preprocessing.	a) Presence of noisy data may degrade the performance. b) Feature selection method degrades the efficiency of classification task	Improved accuracy
P. Carpena, P. A. Bernaola-Galvan, C. Carretero-Campos, A. V. Coronado. (2016). "*Probability distribution of intersymbol distances in random symbolic sequences: Applications to improving detection of keywords in texts and of amino acid clustering in proteins".*	a) Entropic b) Clustering approach	a) Suitable for both long and short texts. b) Reliable obtained results	Median and mode did not give the correct result.	Good clustering results for both short and long texts
Z. Yang, K. Gao, K. Fan, Y. Lai. (2014). *"Sensational headline identification by normalized cross entropy-based metric".*	Shannon entropy	a) Suitable for text with no information known in advance. b) Easy to numerically implement	Median and mode did not give the correct result.	Better results for single document
C. Li, A. Sun, J. Weng, Q. He. (2013).	Hybrid segmentation	High quality tweet segmentation	Manual segmentation is expensive	Improved precision.

"Exploiting hybrid contexts for tweet segmentation".				
J. M. J. Ventura. (2014). ***"Automatic extraction of concepts from texts and applications, Diss"***	Statistical language independent	Good for extracting single and multiword expressions.	Not suitable for long text	Improved precision and recall.
C. W. Wong, R. W. Luk, E. K. Ho. (2005). "Discovering 'title-like2 terms".	a) Decision tree classifier b) Pattern recognition	Easy title determination	a) Not easy to determine the best document size. b) Precision was not significant than recall.	Recall 85% was achieved for title like terms.
D. uttiyapillai, R. Rajeswari, "A method for extracting task-oriented information from biological text sources". (2015).	a) Sensitive text analysis b) Context-based extraction method	a) Category-oriented approach for extraction of task-specific information b) Investigations into recall and precision were carried out.	Not tested on generic data.	a) Food safety is analyzed to prevent future consequences. b) Improved classification accuracy by utilizing optimization constraints. c) Causes of diseases related to low-quality food were identified

Fonte: (SHAIKH, 2018, p. 2593)

Observa-se, no quadro 10, as diferentes técnicas voltadas em cada tipo de análise, suas respectivas vantagens, desvantagens e as principais análises obtidas em cada uma dessas técnicas. Percebe-se que a pesquisa aqui desenvolvida, aproxima-se muito das seguintes abordagens: análise estatística de termos, aprendizagem de máquina, MT e método de extração de termos. Na seção de análise de resultados, será feita uma análise comparativa com a tabela x para verificar as vantagens, desvantagens e principais resultados observados, comparando com outras técnicas já utilizadas em pesquisas no mundo.

2.3.3 Definição sobre o software de Indexação Automática – Maui

O software Maui foi escolhido para ser utilizado nesta pesquisa como o sistema de IA a realizar a tarefa de indexação. Tal sistema é denominado simplesmente de IA de Tópicos

Multiuso. O nome do *software* foi uma homenagem ao herói e semideus da mitologia Maori. (MEDELYAN, 2009). O Software Foi criado em 2009, pela cientista Olena Medelyan, em sua tese de Doutorado na Universidade de Waikato, Nova Zelândia. O *software* utiliza código aberto, é distribuído sob a GNU (*General Public License*) e possui como objetivo principal a indexação por tópicos automatizada. Como afirmaram Silva e Corrêa (2020, p. 3):

> O MAUI (Multi-purpose Automatic Topic Indexing) é um sistema multilíngue de origem neozelandesa que faz uso de um tesauro e algoritmo de aprendizagem de máquina para gerar um modelo a partir de resultados da indexação intelectual, sendo os termos representados por características estatísticas.

O sistema é capaz de realizar as seguintes tarefas de indexação de tópicos: atribuição de termos de um vocabulário controlado; indexação de assunto; indexação de tópico com termos da Wikipédia; extração de palavras-chave; extração de terminologia; marcação automática; extração de terminologia e indexação de tópico semiautomática; extração de palavras-chave com o uso de vocabulário controlado. (SILVA; CORRÊA, 2020, p. 15)

O propósito do sistema é competir com humanos nessa tarefa de indexação, levando-se em consideração sua capacidade superior em rapidez, porém, o interesse maior é de alcançar a mesma qualidade obtidas por humanos. Esse sistema utiliza aprendizagem de máquina supervisionada, no qual o algoritmo tem a capacidade de aprender com exemplos para classificar termos candidatos. O MAUI foi originado do algoritmo de extração de palavras-chave KEA++ (*keyphrase extraction*), algoritmo que se restringia à atribuição de termos para documentos agrícolas. O MAUI implementa novos recursos e um novo classificador melhorando significativamente o desempenho das tarefas de indexação de tópicos em relação ao KEA++. (MEDELYAN, 2009)

Sua proposta implementa uma nova abordagem de atribuição evitando as abordagens baseadas em classificação, isto posto, ele segue a estratégia de extração de palavras-chave de forma supervisionada. Por utilizar a aprendizagem de máquina podem ser aplicados a qualquer domínio e tamanho do vocabulário, desde que um conjunto de documentos indexados manualmente esteja disponível. (MEDELYAN, 2009)

Nesta tese, foi utilizado um vocabulário controlado da área a ser analisada com o propósito de realizar normalização dos termos. De acordo com Silva e Corrêa (2020, p. 18) os principais elementos que caracterizam o sistema podem ser descritos a seguir:

a) Os arquivos utilizados no processamento podem ser: texto completo sem marcações; lista de *stopwords*; vocabulário controlado; conjunto de treinamento envolvendo texto dos

documentos e respectivas palavras-chave, utilizado para treinar o modelo de aprendizagem de máquina.

b) Formato dos arquivos de entrada: txt; vocabulário controlado SKOS.
c) Utiliza, obrigatoriamente, a linguagem documentária para implementar a indexação por atribuição.
d) Possui flexibilidade sendo extensível para outros idiomas e áreas do conhecimento por meio do vocabulário controlado, lista de *stopwords*, *software* radicalizador e conjunto de treinamento.
e) As etapas do processamento iniciam-se pelo treinamento de um modelo de aprendizagem de máquina, em seguida ocorre a geração de termos candidatos à indexação, o cálculo de características para os termos candidatos, logo após a aplicação do modelo de indexação treinado na ponderação e a seleção de termos de indexação.
f) Os tipos de operação sob o texto que podem ocorrer são: realização de análise léxica; remoção de *stopwords;* radicalização dos termos; extração de n-gramas das palavras; adoção do vocabulário controlado.
g) Ocorrência da ponderação dos termos com a seguinte atribuição de parâmetros: Cálculo da frequência de palavras; posição do termo no texto; definição do tamanho do termo; cálculo da probabilidade de ser palavra-chave; definição das relações semânticas.

Baseado no grupamento bibliográfico relatado acima, o próximo capítulo elucidou como ocorreram as etapas metodológicas pretendidas nesta pesquisa. Ressalta-se que os instrumentos escolhidos procuraram validar o percurso metodológico proposto.

3 PROCEDIMENTOS METODOLÓGICOS

Neste capítulo, foram abordadas informações sobre a classificação da pesquisa e sua caracterização. De início, são apontadas as tipologias identificadas no contexto da pesquisa, a partir da descrição das etapas envolvidas; e, por último, são explicitados os procedimentos escolhidos para seu desenvolvimento.

3.1 Caracterização da pesquisa

De acordo com Vergara (2003), a pesquisa metodológica é o estudo que se refere ao desenvolvimento de instrumentos de captação ou manipulação da realidade, portanto, associado a caminhos, formas, maneiras e procedimentos para atingir determinado fim. Sendo assim, o propósito deste trabalho, a saber: propor um percurso metodológico para a formulação de indicadores temáticos de informação científica em bases de dados, com aporte da IA e dos EMI, encaixa-se como pesquisa metodológica. Quanto aos meios, utiliza-se da pesquisa bibliográfica, para apresentar e sistematizar os fundamentos e experiências presentes na literatura – valendo-se de material já publicado e revisado por pares; e da telematizada, buscando informações em bases de dados disponíveis na internet e aplicando recursos algorítmicos para processamento de dados e produção de estatísticas.

Os principais aportes técnicos e teóricos selecionados estão associados à Organização da Informação e do Conhecimento, que discute os princípios da IA, da RI e dos tesauros; os EMI, enfocando a produção de métricas e estatísticas sobre a produção de conhecimento; a aprendizagem de máquina, permitindo a aplicação de modelos computacionais para a conversão de linguagem natural em vocabulários controlados; e o *Business Process Management* (BPM), para a representação de fluxos e processos que podem ser utilizados por outros pesquisadores.

Além disso, a pesquisa tem características empíricas, por basear-se em evidências e resultados obtidos por um conjunto de procedimentos qualiquantitativos, e coaduna com a pesquisa aplicada, utilizada na solução de problemas concretos. Por isso, espera-se abranger todos os aspectos conceituais e metodológicos, inerentes à construção de um percurso, que busque atingir os objetivos apresentados na introdução.

No desiderato de atestar a veracidade dos pressupostos apresentados na introdução, empregou-se a perspectiva hipotético-dedutiva (MARCONI; LAKATOS, 2013, p. 110), em que são identificadas as lacunas nos conhecimentos acerca da formulação das hipóteses por

meio do processo de inferência dedutiva, testando a predição da ocorrência dos fenômenos abrangidos.

3.2 Percurso Metodológico Proposto

Inicialmente, a pesquisa adotou um sistema de IA, que precede a utilização dos EMI, na construção do percurso metodológico. Essa abordagem parte de um pressuposto qualiquantitativo, buscando estabelecer uma relação dinâmica entre o mundo das análises dos dados e da teoria empregada, utilizando-se desta para confrontar ou ratificar o observado nos dados coletados.

Posteriormente, na procura por um enfoque representativo, fez-se uso dos EMI, pois estes consideram as informações numéricas, que foram explicitadas em gráficos e em quadros, servindo-se de técnicas estatísticas para análise dos dados. Do mesmo modo, pretendeu-se realizar análises bibliométricas dos resultados obtidos pela IA, com o intuito de concretizar um levantamento temático durante o desenvolvimento da investigação científica.

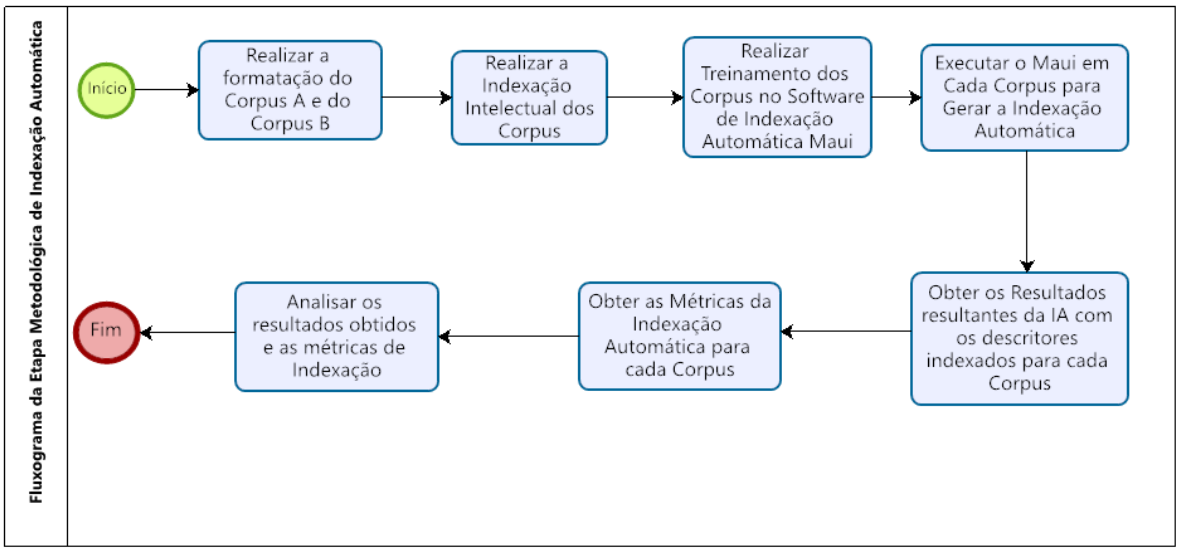
Em seguida, elaboraram-se as possíveis justificativas e os argumentos que puderam gerar a arguição necessária para completa compreensão dos resultados alcançados. Por meio das interpretações teóricas e fenomenológicas, revisitando os objetivos, procura-se atribuir de significado o que foi coletado, permitindo, assim, a tangibilidade do proposto.

Destarte, para atingir a proposição do percurso metodológico, foi necessária a adoção de uma série de etapas que contemplassem os objetivos propostos e pudessem contemplar o problema de pesquisa proposto.

3.2.1 Descrição das etapas do processo de proposição do percurso metodológico

Para compreender o fluxo do processo metodológico, é importante destrinchar as etapas realizadas na pesquisa. Portanto, na figura 6, são observados todos os processos pertinentes ao fluxograma da IA.

Figura 6 – Fluxograma da Indexação Automática



Fonte: Dados da Pesquisa (2021).

Com o intuito de interpretar cada etapa envolvida no percurso, enumerou-se o caminho metodológico que envolve a IA:

1- **Realizar formatação do *corpus* A e do *corpus* B.** Executou-se a formatação dos *corpora* com base nos aspectos técnicos solicitados pelo sistema.

2- **Realizar a indexação intelectual dos *corpora*.** Finalizou-se a indexação intelectual de cada conjunto, verificando a existência de termos que correspondessem às palavras encontradas em cada conjunto de metadados. Para isso, utilizou-se do site (https://vocabularyserver.com/tbci/index.php), a fim de concluir o processo numa linguagem documentária padrão.

3- **Realizar Treinamento do *Corpus* no *software* de IA Maui.** Fez-se o treinamento no *software* Maui, para que ele conseguisse aprender a indexar. Neste processo, só foi utilizado um *corpus* para treinamento, pois os dois pertencem à área da CI, compartilhando, assim, da mesma linguagem.

4- **Executar o Maui em cada *Corpus* para Gerar a IA.** Testou-se o *software* em cada *corpus*, para gerar os termos indexados pelo referido sistema.

5- **Obter os resultados resultantes da IA com os descritores indexados para cada *corpus*.** Verificou-se cada conjunto de termos indexados para seu respectivo *corpus*. Em vista disto, foi possível visualizar o resultado da indexação realizada pelo sistema.

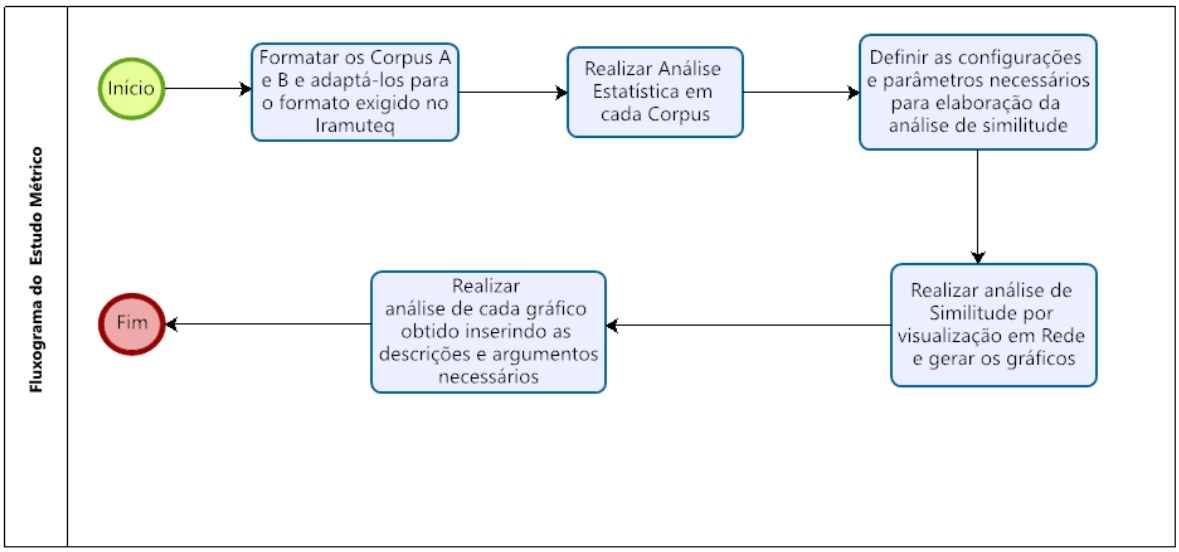
6- **Obter as métricas de IA para cada *corpus***. Os resultados do *software* Maui foram avaliados por meio de medidas que pudessem atestar sua capacidade de indexação em cada *corpus*. Para o cálculo dos índices, adotou-se as seguintes fórmulas: a) Revocação = (Número

de Termos relevantes recuperados/Número total de termos relevantes); b) Precisão = (Número de termos relevantes recuperados/Número total de termos recuperados); c) Medida F = (2 x (Precisão x Revocação) / Precisão + Revocação). A medida F também pode ser considerada a média harmônica ponderada entre a precisão e a revocação.

7- **Analisar os resultados obtidos e as métricas de indexação.** Finalmente, cada conjunto de termos indexados pôde ser analisado, verificando-se os critérios adotados anteriormente. Além disso, foram elaboradas as justificativas necessárias aos resultados obtidos.

Posteriormente, para obter a consolidação e validação do percurso metodológico proposto, construiu-se o fluxograma apresentado na figura 7.

Figura 7 – Fluxograma do Estudo Métrico



Fonte: Dados da pesquisa (2021)

Com o objetivo de compreender cada etapa envolvida no percurso, foram enumeradas e sistematizadas as etapas:

1- **Formatar os *corpora* A e B e adaptá-los para o formato exigido no Iramuteq**. Ambos os *corpora* foram formatados de acordo com as definições técnicas exigidas pelo sistema. Ressalta-se que, para isso, os arquivos utilizados foram alimentados com as palavras-chave e os descritores envolvidos em cada uma das 3 etapas de análise. Para esse fim, foi realizada uma formatação do arquivo de extensão (.txt) correspondente ao *corpus* da pesquisa, no formato padrão exigido pelo IRAMUTEQ. Nesta etapa, o conteúdo dos campos de metadados de cada artigo foi numerado com um padrão de caracteres (*** *Artigo_X, onde X

corresponde a um inteiro único para cada artigo), que permitiu seu processamento. Os arquivos foram salvos com codificação UTF-8 SEM BOOM no bloco de notas.

2- **Realizar análise estatística em cada *corpus***. Adotou-se uma análise de zipf para demonstrar a aplicação da Lei de Zipf, observando o grau de dispersão das palavras em ambos os *corpora*. Ressalta-se que lei foi utilizada no contexto específico desta tese, analisando metadados previamente selecionados, portanto, não considera o texto completo nesse tipo de análise. Assim, buscou-se comparar a frequência e a ocorrência das palavras, nesse contexto.

3- **Definir as configurações e parâmetros necessários para a elaboração da análise de similitude.** Com o objetivo de alcançar uma melhor visualização dos agrupamentos temáticos de termos, e obter os núcleos semânticos, determinaram-se as configurações a seguir: a) Análise de Similitude -> propriedades -> zerar as palavras-vazias; b) Definição das palavras que iriam ser utilizadas nas análises; c) Configurações gráficas, (Escore Adotado -> coocorrência para identificar agrupamentos de palavras que ocorrem de maneira próxima; Apresentação -> Método de visualização, *Kamada Kawai*, para permitir que as palavras permanecesse mais destacadas no gráfico; Tipos de gráfico -> foi adotado o modo Dinâmico para alterar as posições e optou-se por Árvore Máxima; Selecionaram-se as opções Comunidades e Halo, e a configuração edge.betweenness.community, para destacar os agrupamentos; Dentro dos Ajustes Gráficos, escolheu-se o tamanho do vértice proporcional à frequência);

4- **Realizar análise de similitude por visualização em rede e gerar os gráficos.** Após a devida configuração e definição dos parâmetros para geração dos gráficos, foi possível identificar a visualização da informação por meio dos gráficos de similitude. Dessa forma, as visualizações auxiliaram na compreensão semântica e temática, favorecendo a compreensão. A construção da análise de similitude baseia-se na teoria dos grafos, possibilitando identificar as coocorrências entre as palavras. Esse resultado traz indicações de conexidade entre as palavras, auxiliando na identificação da estrutura de um *corpus* textual, distinguindo as partes comuns e as especificidades das variáveis descritas. (FERREIRA; CORRÊA, 2018).

5- **Realizar análise de cada gráfico obtido, inserindo as descrições e argumentos necessários**. Analisaram-se os resultados, contrapondo cada um dos agrupamentos temáticos a fim de atestar a validade do percurso. Portanto, examinaram-se os resultados obtidos e elaboraram-se as justificativas necessárias em cada *corpus*; além disso, foram construídas inferências para ampliá-la de forma crítica.

3.3 Compilação dos *corpora*

Necessita-se delimitar o ambiente para geração e análise dos dados. Neste estudo, foram selecionados dois *corpora* a serem analisados, que constituem um conjunto de documentos bibliográficos já publicados, mais especificamente:

a) O *corpus* A corresponde aos documentos selecionados por Souza (2005, p. 72) em sua Tese de Doutorado. O autor selecionou 60 artigos publicados em duas revistas da época: 29 da Datagramazero e 31 da Ciência da Informação. Esses artigos foram escolhidos por sua importância e qualidade reconhecida pelo Qualis Capes naquele momento. Esse *corpus* está identificado na Tabela 5 do anexo A da tese de Souza (2005). Ressalta-se que foi reutilizado o *corpus* compilado por Silva, Corrêa e Gil-Leiva (2020), alterando os .txt para incluir título, resumo, e as palavras-chave dos autores. Os .key não foram alterados, mantendo-se a indexação intelectual. O processo de modificação dos .txt auxilia no treinamento do modelo de indexação, levando em conta os termos atribuídos pelos indexadores a cada documento do conjunto do treinamento (SILVA; CORRÊA; GIL-LEIVA, 2020, p. 15)
b) O *corpus* B corresponde aos 82 artigos utilizados por Ferreira e Corrêa (2018). Os autores utilizaram trabalhos que foram publicados em periódicos e estão indexados na Brapci (Base de Dados em Ciência da Informação), por possuírem "Biblioteca digital" no campo das palavras-chave e serem identificados como artigos de periódicos brasileiros. Esse *corpus* está identificado na Tabela 6 do Apêndice B. Destaca-se que, para a obtenção dos arquivos .key, foram utilizados arquivos .txt, com o conjunto de metadados (título, resumo e palavras-chave) correspondente a cada documento. A partir disso, foi realizada a indexação intelectual, utilizando como formato padrão o TBCI[35]. Neste processo, foram observados os metadados presentes nos arquivos .txt, atribuindo termos genéricos com base na linguagem controlada. Assim, foi possível formatar os arquivos .key, utilizando os descritores observados no Tesauro.

A escolha dessa coleção de documentos foi definida pelo fato de estes já terem sido analisados por outros autores da Ciência da Informação. Como o propósito desta pesquisa é

[35] De acordo com Silva, Corrêa e Gil-Leiva (2020, p.15), O tesauro compreende cerca de 1.800 termos os quais, em sua maioria, possui versão em língua inglesa e espanhola e são complementados por definições. Possui a finalidade de assumir um papel essencial na recuperação da informação na área de Ciência da Informação no Brasil e em outros países. Com relação ao uso, o tesauro é direcionado aos indexadores, professores, pesquisadores e qualquer profissional da informação.

propor um percurso metodológico, a delimitação da área do conhecimento torna possível esse processo, e a escolha desses dados possibilita observar de forma criteriosa os resultados obtidos.

3.4 Procedimentos necessários para a compilação de dados

A primeira coleta de dados foi realizada por meio da análise dos metadados temáticos[36], além do texto completo, dos 60 artigos para o *corpus* A, e 82 metadados temáticos o *corpus* B. Vide Anexo A.

3.4.1 Análise de Dados por meio da Ferramenta de Indexação Automática – Maui

Foi adotado um *software* para a geração dos resultados pretendidos com a IA, o MAUI. Esse sistema foi escolhido devido a sua possibilidade de processamento do texto completo ou resumo, sem a necessidade de marcação. Além disso, a indexação por atribuição, empregada pelo *software*, permite a utilização de um vocabulário controlado, o que possibilita maior eficiência nos resultados.

Nesta etapa, o *software* Maui foi baixado (maui-pt-master) e extraído numa pasta de escolha do pesquisador. A partir disso, obteve-se acesso à subpasta, *data*. Dentro dessa subpasta, foram encontradas outras, que fazem parte do programa, e os arquivos com extensão .bat, responsáveis por toda a execução. Na subpasta *docs*, estavam os documentos/arquivos em que o *software* baseou sua aprendizagem de máquina. Nela, identificaram-se as pastas para treinamento e teste, com o objetivo executar o programa. Numa pasta de treinamento, por exemplo, identificam-se os artigos (textos completos) ou metadados (títulos, palavras-chave, resumos), nos quais foram baseadas as indexações intelectuais. Nesse caso, cada arquivo de texto foi associado a um arquivo de extensão .key, que contém as indexações intelectuais – geralmente a indexação intelectual é determinada pelas palavras-chave atribuídas pelo autor ou descritores pelos indexadores.

Basicamente, o sistema trabalha no formato de entrada/saída, utilizando-se de um algoritmo de aprendizagem de máquina que realiza o processo. Logo, a entrada pode ser um

[36] Para esta pesquisa, deverão ser considerados os metadados temáticos, como sendo: títulos, resumos e palavras-chave.

artigo científico, por exemplo, e a saída ideal[37], os termos que foram encontrados pelos indexadores intelectuais. Voltando para a subpasta *data*, é possível reconhecer a pasta *models*, que comporta o modelo de aprendizagem de máquina, gerado pelo *software*. Esse arquivo é armazena a memória de aprendizagem do sistema. Na subpasta *vocabulary*, observam-se os vocabulários controlados, utilizados pelo sistema para atribuir os termos corretos com base nessa nomenclatura. Nesta pesquisa, como foram empregados documentos da área de CI, optou-se pela utilização pelo TBCI[38].

3.4.2 Treinamento e Execução do Software Maui

Após ter sido ser realizada a devida indexação intelectual dos documentos com base no TBCI, os arquivos .key foram gerados e associados a cada artigo/documento que vai ser processado pelo sistema. Em seguida, compilou-se o arquivo .bat de treinamento, chamado *train*, que continha todos os parâmetros necessários para o treinamento do sistema; o nome do arquivo foi salvo como modelo de treinamento, na pasta *models*, e a definição do vocabulário controlado, como linguagem de indexação, que baseou o seu treinamento, conforme visualizado na figura 3 do Apêndice A. Ao executar o arquivo *train*, o *software* processou o aprendizado e construiu um modelo de treinamento na pasta *models*.

Em seguida, testou-se o modelo de aprendizagem criado, executando o arquivo test.bat.. Nesse caso, o MAUI indexou cada documento e comparou sua indexação com a realizada pelos indexadores humanos. Ao final da execução, ele apresentou as medidas de precisão, revocação e medida-f, atestando sua capacidade de indexação do sistema, situação observada na figura 4 do Apêndice A. Na última etapa, ao ter sido executado o arquivo run.bat, um documento selecionado indexou automaticamente, com base nos parâmetros determinados, e apresentou, como saída, as palavras-chave e seus índices de probabilidade, etapa esta identificada na figura 5 do Apêndice A.

[6] Num processo de indexação automática é esperado que o sistema obtenha resultado o mais próximo possível de um indexador intelectual. Essa é a saída ideal.

[38] Disponível em: Tesauro Brasileiro de Ciência da Informação (http://www.uel.br/revistas/informacao/tbci/vocab/). De acordo com Vogel e Jabala (2019) o tesauro é um instrumento de controle de vocabulário com fins de organização e recuperação da informação. Neste caso, o Tesauro utilizado nesta Tese contempla à área da Ciência da Informação pois os documentos utilizados na indexação pertencem a esta área.

3.5 Análise de Dados por meio da Ferramenta dos Estudos Métricos – Iramuteq

A análise dos dados ocorreu por meio das categorias: variáveis quantificáveis dos resultados obtidos após a utilização do *software* de IA; variáveis qualiquantitativas dos *softwares* utilizados na indexação e as variáveis quantificáveis dos metadados dos artigos publicados sobre temática previamente escolhida. Inicialmente, adotou-se as seguintes análises: aplicação da lei de Zipf, frequência e ocorrência de palavras, análise de similitude, frequência das palavras-chaves compostas. Posteriormente, buscou-se a comparação dos resultados obtidos buscando a construção de inferências e deduções a partir dos dados.

Sobre o *software* Iramuteq, de acordo com Camargo e Justo (2013 apud FERREIRA; CORRÊA, 2018), é um software gratuito, desenvolvido, originalmente, no idioma francês, pelo cientista Pierre Ratinaud, em 2009, como ferramenta de organização de dados. Ele ancora-se no ambiente estatístico do *software* R (www.r-project.org) e na linguagem python. Permite a viabilização de diferentes tipos de análise de dados textuais, como lexicografia básica, lematização, cálculo de frequência de palavras, análises multivariadas, análise pós-fatorial e análise de similitude. O intuito de utilizar esse *software* é possibilitar uma vasta análise estatística que possa também oferecer recursos visuais que facilitem o estudo.

3.5.1 Descrição do funcionamento e Configuração do Software Iramuteq

Após o devido download do software Iramuteq no site (http://www.iramuteq.org/), buscou-se baixar a versão mais recente do software R no site (https://vps.fmvz.usp.br/CRAN/), assim, foi possível gerar as análises com base nessa ferramenta. Em seguida, coletaram-se os termos indexados pelo MAUI inseridos num arquivo de extensão (.txt) que foram analisados. Nessa etapa, os termos representativos indexados, serviram como dados necessários para as análises propostas, que foram descritas nas fases a seguir:

Fase 1- Após a realização do processamento textual foram geradas análises estatísticas dos termos, identificando formas, números de ocorrências e número de formas das palavras, frequência dos termos, termos que foram citados uma única vez (hápax) e a referida classe gramatical de cada palavra, (situação que pode ser observada no exemplo da figura 6 no Apêndice B);

Fase 2- Nesta subetapa, analisaram-se as diferentes frequências absolutas dos termos mais frequentes. O objetivo é identificar as palavras mais significativas, utilizadas de forma repetitiva para compor o conjunto de representação semântica do texto.

Fase 3- Posteriormente, foram gerados gráficos de similitude das palavras presentes nos resumos dos artigos completos e metadados das teses utilizadas na amostra. Com esses gráficos foi possível comparar, com as outras análises subsequentes, buscando identificar um determinado grau de proximidade entre os termos indexados pelo MAUI e entre a árvore de similitude. Essa análise comparativa buscou identificar os núcleos centrais dos termos, suas ramificações e o graus de conexidade. O objetivo principal desta etapa foi perceber se os termos presentes nos núcleos centrais foram indexados pelo *software*.

Fase 4- Logo depois, realizou-se as análises de frequência de palavras-chave dos autores, dos descritores da indexação intelectual e dos descritores da indexação automática, em ambos os *corpora*. Nessas análises demonstraram-se um conjunto de palavras ordenadas pelo índice de frequência, num gráfico de barras horizontal, elaborado no Microsoft Excel 365. O principal objetivo é o de identificar quais termos possuem maior importância no *corpus* textual a partir dos descritores identificados. Essa fase buscou identificar a frequência dos termos, assim como os termos mais relevantes dos documentos.

3.6 Análise do percurso metodológico

Por fim, o percurso metodológico foi analisado por meio da verificação dos resultados obtidos pelas técnicas adotadas. Para tanto, testaram-se, de acordo com suas funcionalidades, em dois *corpora* de dados específicos, verificando o nível de eficiência e a aplicabilidade das ferramentas. A eficiência do percurso metodológico foi inicialmente validada pelo índice de palavras-chave recuperadas com o *software* de indexação Maui; em seguida, pretendeu-se validar o percurso pelas formas de visualização da informação propostas por cada técnica do *software* Iramuteq. Ao final, procurou-se levantar pressupostos que pudessem validar o percurso metodológico, como pode ser observado no quadro a seguir:

Quadro 11 – Pressupostos a serem verificados na validação do percurso

Pressupostos da Pesquisa	Verificação dos Pressupostos
1- A dispersão terminológica das palavras-chave informadas nos registros bibliográficos é um fator limitante da realização de estudos métricos sobre as temáticas.	**Verificou-se o índice de dispersão terminológica na atribuição das palavras-chave, comparando os termos atribuídos pelo autor e as palavras indexadas pelo software Maui.**
2- Os sistemas de IA podem ser utilizados como instrumentos na proposição de um percurso metodológico para a representação de temáticas	**Fez-se a avaliação para verificar se os resultados obtidos pelo Maui, após a IA, oferecem segurança para gerar bons resultados nas visualizações com o software Iramuteq.**
3- A IA permite maior rapidez no que tange à atribuição de termos em documentos, com isso, ela torna-se vantajosa, pois os custos envolvidos em contratação e treinamento de pessoas para realizá-la manualmente.	**Avaliou o grau de benefícios e performance da IA em relação à manual.**
4- O processo de IA associado aos EMI auxilia na validação da coleta de vários IT, quando se trabalha com grandes volumes de textos, permitindo a identificação de descritores e conceitos.	**Verificou se as potencialidades da aplicação conjunta dos EMI e da IA, sobretudo, no que concerne à produção de métodos de visualização de IT.**

Fonte: o autor (2021).

O quadro 11 apresenta os pressupostos que deverão ser verificadas após a finalização dos resultados e análise de desempenho de ambos os sistemas. Durante a validação do percurso metodológico, deverão ser descritas as arguições necessárias a respeito de cada resultado obtido e como isso influenciou no processo de escolha de cada ferramenta. Esses argumentos serão descritos no capítulo de análise dos resultados e considerações finais.

A seguir, apresenta-se o quadro 12, no qual podem ser visualizados os objetivos específicos e suas principais ações para sua consecução.

Quadro 12 – Objetivos Específicos verificados na pesquisa

Objetivos específicos da Pesquisa	Ação para atingir o Objetivo
1- Identificar os pontos de integração teórica e metodológica entre a IA e os EMI.	Para isto, fez-se uma revisão de literatura sobre estas possibilidades de integração, selecionando, especificamente, trabalhos que desenvolveram experiências nesta linha.
2- Delinear a trajetória metodológica de formulação de IT de informação científica	A trajetória metodológica foi composta com base na junção entre os EMI e a IA, acionando-se os aportes do BPM para registro dos fluxos, visando facilitar a replicação por outros pesquisadores.
3- Implementar as análises e processamentos nos corpora definidos na pesquisa, validando os resultados obtidos no percurso metodológico.	Comparação da IA e manual com base nos IT produzidos pela utilização do MAUI e IRAMUTEQ.

Fonte: o autor (2021).

O quadro 12 destaca os objetivos específicos estabelecidos e as principais ações para seu atingimento. Ressalta-se que o conjunto destes objetivos converge para alcançar o objetivo geral: propor um modelo metodológico para a formulação de indicadores temáticos de informação científica, a partir da integração teórica e metodológica entre a IA e os EMI.

Com o intuito de oportunizar continuidade a presente tese, na seção a seguir, é apresentada a análise de resultados com o respectivo processo de validação do percurso metodológico.

4 ANÁLISE DOS RESULTADOS

Os resultados estão estruturados em três tópicos principais, o primeiro, 4.1, apresenta a os produtos da IA, enfatizando as etapas do fluxograma decorrente do percurso metodológico para a geração de métricas de indexação dos *corpora* A e B; o 4.2 expressa os resultados obtidos pelo software IRAMUTEQ e a aplicação da Lei de Zipf para a observação do comportamento das palavras-chave numa estrutura de dispersão; e o 4.3 contempla as configurações e parâmetros para a análise de similitude, a visualização em rede dos termos com suas relações de coocorrência e a análise de grupamentos envolvendo as indexações automática e manual.

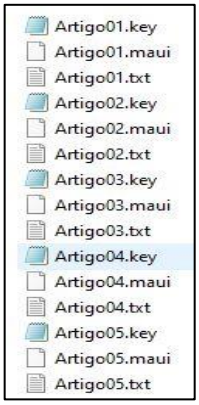
4.1 Análise das Etapas da Indexação Automática

Abaixo, encontram-se as etapas realizadas para a elaboração do processo de IA. A natureza das atividades descritas está ligada a processos de coleta, conformação e processamento de dados com vistas à obtenção de métricas de IA. Para isto, aplicações da Organização da Informação e da Ciência de Dados foram fundamentais. Assim, seguem as etapas que compõem o fluxo construído.

Etapa 1 do fluxograma – Realizar a formatação do *corpus* A e do *corpus* B

Para esta etapa foi necessária a seleção dos *corpora* para o início das análises. Sobre isto, entende-se que os dados devem pertencer a uma mesma área do conhecimento ou domínio. Ou seja, se o primeiro *corpus* escolhido for da área de Microbiologia, necessariamente, todos os *corpora* seguintes deverão ser da mesma área. Para a análise consolidada neste estudo, compreende-se que os dois *corpora* estão adequados por pertencerem a uma mesma área (CI) e por terem sido publicados em revistas com *Qualis* Capes, estabelecendo-se características comuns, fruto do pertencimento a um mesmo domínio. Dessa maneira, ambos os *corpora* foram formatados, definindo-se os títulos, resumos e palavras como os elementos fundamentais por se tratar de partículas semânticas essenciais para a realização da IA por atribuição. Assim, estes metadados foram postos em um arquivo de extensão .txt, como pode-se observar figura 8, onde estão sistematizados em uma pasta, que serviu de entrada para o *software* Maui.

Figura 8 – Arquivos Formatados para Análise

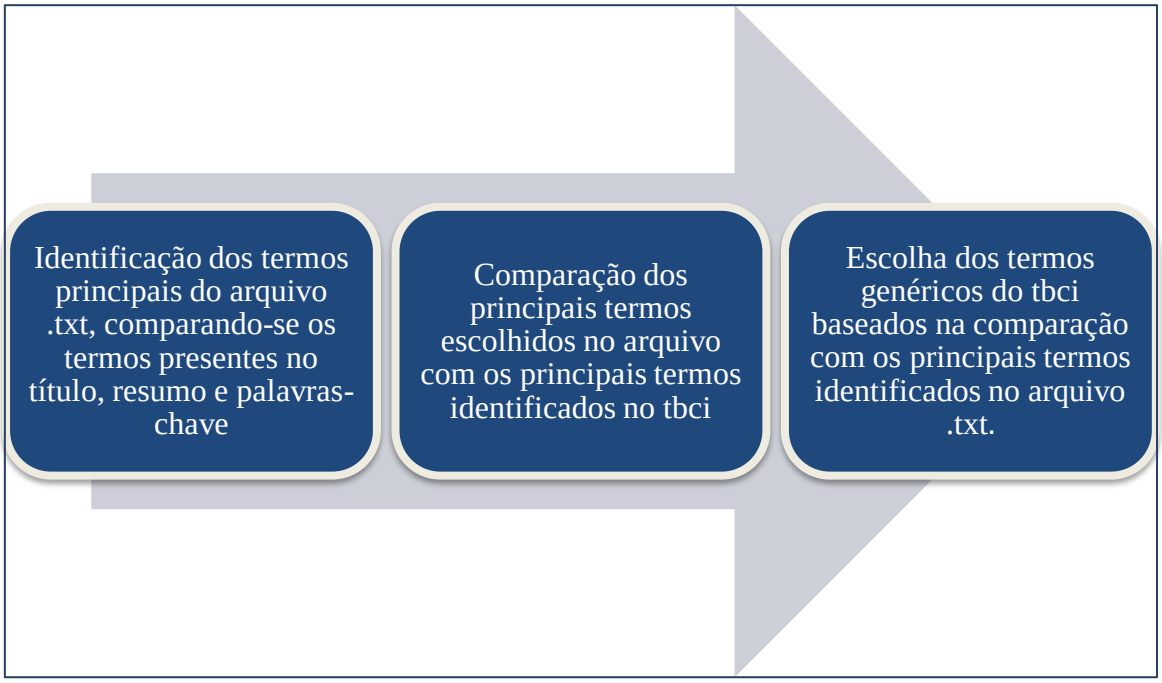


Fonte: dados da pesquisa (2021).

Etapa 2 do fluxograma – Realizar a indexação intelectual dos *corpora*

Para esta etapa foi necessário realizar uma indexação intelectual, verificando a existência de termos que correspondessem às palavras encontradas em cada conjunto de metadados. Essa verificação foi feita no TBCI. Na figura 9, é possível verificar o fluxo adotado para a indexação intelectual.

Figura 9 – Fluxo de Etapas Adotadas na Indexação Intelectual

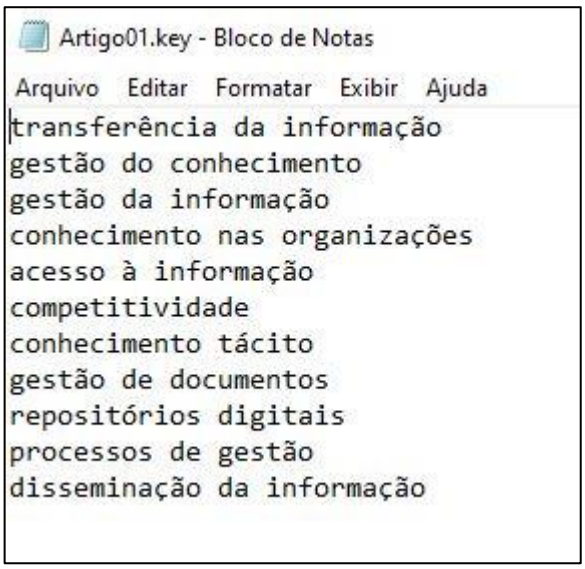


Fonte: dados da pesquisa (2021).

Na figura 10, observa-se o interior de um arquivo de extensão .key. Este arquivo foi elaborado de acordo com o procedimento previamente definido na Indexação manual e

intelectual por atribuição de termos da linguagem controlada. A figura 10 exemplifica quais termos foram adicionados para o Artigo 01 do *Corpus* A.

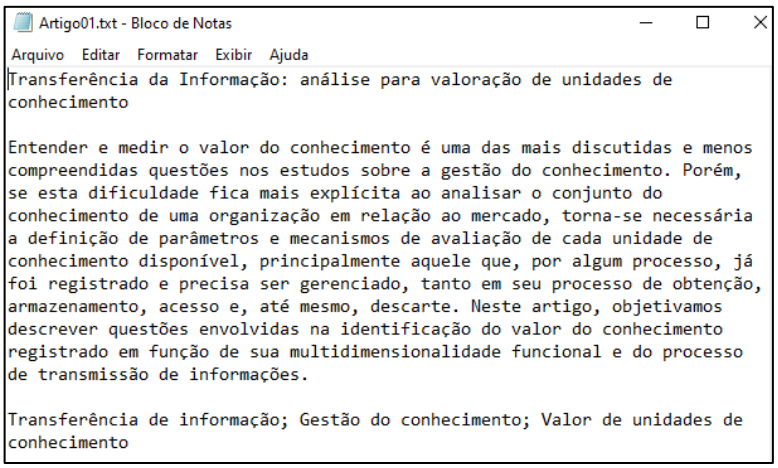
Figura 10 – Exemplo de Arquivo com extensão .key formatado



Fonte: dados da pesquisa (2021).

Na figura 11, identificam-se as estruturas de metadados textuais retirados dos artigos que serviram como instrumento de aprendizado de máquina pelo sistema de IA. O título, o resumo e as palavras-chave foram as estruturas textuais escolhidas que formaram os três elementos principais utilizados pelo sistema como parâmetro.

Figura 11 – Exemplo de arquivo com extensão .txt formatado



Fonte: dados da pesquisa (2021).

Etapa 3 do fluxograma - Realizar Treinamento do *Corpus* no *software* de IA Maui.

Esta etapa é justificada por uma razão técnica, o *software* necessita identificar que padrão terminológico deve ser utilizado para indexar os documentos. A IA só ocorre com o aprendizado da linguagem controlada, neste caso, o TBCI. Entende-se que o vocabulário controlado é fundamental para evitar a dispersão dos termos, pois conforme indica Siqueira (2011), esta linguagem intervém na organização e na articulação dos pontos de acesso, a fim de minimizar os principais problemas de um sistema informacional, que na visão deste autor, são a duplicação e a dispersão informacional.

Nessa circunstância, o treinamento do *software* foi realizado, de acordo com a etapa desenvolvida na subseção 3.4.1. Os parâmetros utilizados levaram em consideração as exigências delimitadas pelo formato de arquivo previamente definido. E todo o processo de treinamento foi feito, seguindo o protocolo de atuação previamente estabelecido.

Etapa 4 do fluxograma - Executar o Maui em cada *Corpus* para Gerar a IA

Nesta etapa, o arquivo "teste.bat" foi executado em cada *corpus*, após a devida etapa de indexação e de treinamento. Neste ponto, o sistema executou a IA dos *Corpora*, resultando nos termos extraídos para análise, assim como, na definição das medidas de consistência, precisão, revocação e medida-f. Ao final, considera-se que o Maui foi bem-sucedido nesta etapa, realizando a indexação dentro dos padrões esperados.

Etapa 5 do fluxograma - Obter os resultados resultantes da IA com os descritores indexados para cada *corpus*

Nesta etapa do processo, o Maui gerou como saída um arquivo de extensão .txt com as métricas encontradas e os respectivos descritores de cada artigo indexado. Os descritores podem ser visualizados nos Apêndices A e B desta tese.

Etapa 6 do fluxograma – Obter as métricas de IA para cada *corpus*

Nesta etapa, foram levantadas as principais métricas de indexação associadas ao percurso metodológico que envolve a Indexação Automática no software Maui. O principal propósito desta fase foi o de identificar os níveis de eficiência e eficácia da indexação, analisando se os termos apresentados, estavam alinhados com as palavras e conceitos previamente escolhidos pelos autores e indexadores manuais.

Aqui, foram obtidos os resultados das métricas de IA de ambos os *corpora*, que podem ser identificados na tabela. Na tabela 2, é possível identificar os principais resultados das métricas obtidos por meio da IA do *Corpus* A. Os resultados foram alcançados pelo Maui e parametrizados no *software* Microsoft Excel 365.

Tabela 2 - Resultados do Maui no *Corpus* A

DOCUMENTO	TERMOS EXTRAÍDOS DO MAUI	TERMOS INDEXADOS MANUALMENTE	NÚMERO DE TERMOS COMUNS	CONSISTÊNCIA	PRECISÃO	REVOCAÇÃO	MEDIDA-F
ARTIGO01	7	11	3	0,2	0,4286	0,2727	0,3333
ARTIGO02	6	10	5	0,4545	0,8333	0,5	0,625
ARTIGO03	6	10	2	0,1429	0,3333	0,2	0,25
ARTIGO04	10	10	2	0,1111	0,2	0,2	0,2
ARTIGO05	9	10	4	0,2667	0,4444	0,4	0,4211
ARTIGO06	10	10	6	0,4286	0,6	0,6	0,6
ARTIGO07	9	11	6	0,4286	0,6667	0,5455	0,6
ARTIGO08	10	14	3	0,1429	0,3	0,2143	0,25
ARTIGO09	10	12	5	0,2941	0,5	0,4167	0,4545
ARTIGO10	4	12	3	0,2308	0,75	0,25	0,375
ARTIGO11	7	10	5	0,4167	0,7143	0,5	0,5882
ARTIGO12	10	10	7	0,5385	0,7	0,7	0,7
ARTIGO13	8	13	5	0,3125	0,625	0,3846	0,4762
ARTIGO14	10	12	5	0,2941	0,5	0,4167	0,4545
ARTIGO15	6	11	3	0,2143	0,5	0,2727	0,3529
ARTIGO16	8	11	6	0,4615	0,75	0,5455	0,6316
ARTIGO17	10	11	4	0,2353	0,4	0,3636	0,381
ARTIGO18	2	15	2	0,1333	1	0,1333	0,2353
ARTIGO19	10	13	4	0,2105	0,4	0,3077	0,3478
ARTIGO20	6	11	3	0,2143	0,5	0,2727	0,3529
ARTIGO21	7	9	5	0,4545	0,7143	0,5556	0,625
ARTIGO22	10	7	5	0,4167	0,5	0,7143	0,5882

ARTIGO23	4	11	4	0,3636	1	0,3636	0,5333
ARTIGO24	5	10	2	0,1538	0,4	0,2	0,2667
ARTIGO25	8	13	6	0,4	0,75	0,4615	0,5714
ARTIGO26	7	9	2	0,1429	0,2857	0,2222	0,25
ARTIGO27	9	9	4	0,2857	0,4444	0,4444	0,4444
ARTIGO28	10	9	5	0,3571	0,5	0,5556	0,5263
ARTIGO29	8	7	4	0,3636	0,5	0,5714	0,5333
ARTIGO30	6	14	2	0,1111	0,3333	0,1429	0,2
ARTIGO31	10	9	5	0,3571	0,5	0,5556	0,5263
ARTIGO32	10	8	5	0,3846	0,5	0,625	0,5556
ARTIGO33	10	13	5	0,2778	0,5	0,3846	0,4348
ARTIGO34	8	13	4	0,2353	0,5	0,3077	0,381
ARTIGO35	10	7	6	0,5455	0,6	0,8571	0,7059
ARTIGO36	10	7	3	0,2143	0,3	0,4286	0,3529
ARTIGO37	10	12	7	0,4667	0,7	0,5833	0,6364
ARTIGO38	10	14	5	0,2632	0,5	0,3571	0,4167
ARTIGO39	10	7	5	0,4167	0,5	0,7143	0,5882
ARTIGO40	10	12	5	0,2941	0,5	0,4167	0,4545
ARTIGO41	10	11	3	0,1667	0,3	0,2727	0,2857
ARTIGO42	10	12	9	0,6923	0,9	0,75	0,8182
ARTIGO43	5	8	4	0,4444	0,8	0,5	0,6154
ARTIGO44	8	11	4	0,2667	0,5	0,3636	0,4211
ARTIGO45	10	8	3	0,2	0,3	0,375	0,3333
ARTIGO46	10	7	5	0,4167	0,5	0,7143	0,5882
ARTIGO47	7	12	6	0,4615	0,8571	0,5	0,6316
ARTIGO48	8	9	5	0,4167	0,625	0,5556	0,5882
ARTIGO49	10	11	4	0,2353	0,4	0,3636	0,381
ARTIGO50	7	9	4	0,3333	0,5714	0,4444	0,5
ARTIGO51	4	11	3	0,25	0,75	0,2727	0,4
ARTIGO52	10	9	6	0,4615	0,6	0,6667	0,6316

ARTIGO53	9	8	3	0,2143	0,3333	0,375	0,3529
ARTIGO54	10	10	5	0,3333	0,5	0,5	0,5
ARTIGO55	7	9	3	0,2308	0,4286	0,3333	0,375
ARTIGO56	10	10	7	0,5385	0,7	0,7	0,7
ARTIGO57	4	9	3	0,3	0,75	0,3333	0,4615
ARTIGO58	6	7	5	0,625	0,8333	0,7143	0,7692
ARTIGO59	10	10	7	0,5385	0,7	0,7	0,7
ARTIGO60	10	13	8	0,5333	0,8	0,6154	0,6957

Fonte: dados da pesquisa (2021).

Na tabela 5 são identificados os principais resultados da análise de IA por atribuição do *Corpus* A. Os resultados apontados indicam dois tipos de avaliação adotados: intrínseca (mensura o grau de consistência) e extrínseca (mede o grau de revocação, precisão e medida F).

Etapa 7 do fluxograma - Analisar os resultados obtidos e as métricas de indexação

De acordo com Narukawa (2011, p. 72), a avaliação intrínseca pode ser definida como o conjunto de tarefas centradas no resultado da qualidade da indexação, avaliando a consistência da indexação e o seu grau de concordância na representação da informação de um documento. Enquanto a avaliação extrínseca procura avaliar de forma quantitativa os índices de revocação, precisão e medida f na indexação automática.

Portanto, pode-se afirmar que os termos obtidos por meio da IA, atribuídos pelo Maui, apresentam os valores descritos acima, para o cálculo dos índices de precisão, revocação e medida F. Consequentemente, os resultados da tabela contemplam os seguintes campos: Termos Extraídos do Maui, Termos Indexados Manualmente, Número de Termos Comuns, Consistência, Precisão, Revocação e Medida F. A partir desses resultados, observam-se um mínimo de sete palavras indexadas e um máximo de 15 termos atribuídos pela Indexação Manual, enquanto o Maui atribuiu um mínimo de quatro palavras indexadas e um máximo de 10 termos indexados automaticamente. Esses valores resultam numa média de 10 descritores atribuídos por documento para a Indexação Manual, de um total de 621 descritores atribuídos, e uma média de oito descritores atribuídos por documento, de um total de 495 descritores atribuídos para a Indexação do Maui.

Outro ponto relevante, diz respeito à consistência. Percebe-se uma média de 33% nos índices de Consistência da Indexação, variando entre 11% e 69%, portanto, os resultados da

Indexação no *Corpus* A estão dentro do parâmetro esperado para a média de consistência obtida. Com relação à revocação, foi identificada uma média de 45% no índice indicando um desempenho esperado para o conjunto analisado.

Já em relação à Medida F, foi obtido um índice de 48%. Por ser uma medida que avalia a média harmônica ponderada, entre o índice de precisão (P) e o índice de revocação (R), ela identifica o índice de termos relevantes recuperados. Nesse caso, se nenhum termo relevante fosse recuperado, seria considerado o valor de zero, já se todos os termos recuperados fossem relevantes, assumir-se-ia o valor de um. Para os valores identificados no *Corpus* A, a medida f considerada é de 48%. Desse modo, quase metade dos termos recuperados foram considerados relevantes de acordo com as medidas de precisão e revocação consideradas.

Com o intuito de facilitar a compreensão dos principais valores identificados na análise da Tabela 3, foi elabora a Tabela 3, a seguir:

Tabela 3 - Média das Métricas Obtidas pelo Maui no *Corpus* A

	Número de termos do Maui	**Número de termos da Indexação Manual**	**Nº de termos comuns com a Indexação Manual**	**Consistência**	**Precisão**	**Revocação**	**Medida F**
Mínimo	4	7	2	11%	20%	13%	20%
Máximo	10	15	9	54%	90%	85%	81%
Média	8	10	4	33%	56%	45%	48%
Desvio Padrão	2,13	2,082	1,581	0,136	0,184	0,174	0,152

Fonte: Adaptado de Silva (2020).

Os resultados observados na Tabela 3 apresentam uma média de 33% no índice de consistência, 56% no de precisão e 45% no de revocação. Portanto, o desempenho obtido pela IA está na faixa de 33% a 56%. Desse modo, apesar das diferenças no domínio dos documentos, dos tesauros, e no cálculo das métricas de Revocação e Precisão, os valores obtidos na presente tese para as métricas de qualidade na indexação automática por atribuição são próximos aos encontrados por Narukawa, Gil Leiva e Fujita (2009) para o domínio de Odontologia, a saber: índice médio de Consistência de 23,25%, índice médio de Precisão de 40,92%, e índice médio de Revocação de 35,72%) (BANDIM; CORREA, 2017, p. 9).

Com o *software* de IA SISA, os autores obtiveram os seguintes índices: média de 15% no índice de consistência, com uma variação de um mínimo de 0% a um máximo de 42%.

Quanto aos índices de Revocação, Precisão e Medida F, seguem os seguintes resultados: Precisão média de 20%; Revocação média de 42% e Medida F média de 25% (BANDIM; CORREA, 2017).

Sendo assim, as medidas obtidas pelo Maui correspondem aos valores esperados, apesar da existência de alguns fatores que podem influenciar, negativamente, o processo da IA, como foi apontado por Silva e Corrêa (2020).

Outra observação importante, que foi elaborada com base nos resultados obtidos, e que já foi apontada pelos autores supracitados, diz respeito aos fatores intervenientes que dificultam o aumento da precisão dos processos realizados. Questões de ordem política na atribuição dos termos, de ordem morfológica, linguística e semântica podem dificultar a escolha adequada dos termos mais relevantes de indexação, interferindo na construção das métricas devido a questões subjetivas que permeiam o processo de indexação.

Posteriormente, para dar prosseguimento às análises propostas, foi elaborada a tabela 7. Nela é possível identificar os principais resultados das métricas obtidos por meio da IA do *Corpus* B. Os resultados foram processados e sistematizados pelo Maui e parametrizados no *software* Microsoft Excel 365.

Tabela 4 - Resultados Extraídos do Maui no *Corpus* B

DOCUMENTO	TERMOS EXTRAÍDOS DO MAUI	TERMOS INDEXADOS MANUALMENTE	NÚMERO DE TERMOS COMUNS	CONSISTÊNCIA	PRECISÃO	REVOCAÇÃO	MEDIDA-F
ARTIGO01	5	3	3	0,6	0,6	1	0,75
ARTIGO02	10	3	3	0,3	0,3	1	0,4615
ARTIGO03	10	7	5	0,4167	0,5	0,7143	0,5882
ARTIGO04	10	3	3	0,3	0,3	1	0,4615
ARTIGO05	9	6	6	0,6667	0,6667	1	0,8
ARTIGO06	10	6	4	0,3333	0,4	0,6667	0,5
ARTIGO07	10	5	4	0,3636	0,4	0,8	0,5333
ARTIGO08	7	3	3	0,4286	0,4286	1	0,6
ARTIGO09	10	7	5	0,4167	0,5	0,7143	0,5882
ARTIGO10	10	8	7	0,6364	0,7	0,875	0,7778
ARTIGO11	10	4	3	0,2727	0,3	0,75	0,4286
ARTIGO12	10	7	6	0,5455	0,6	0,8571	0,7059
ARTIGO13	6	5	4	0,5714	0,6667	0,8	0,7273

ARTIGO14	10	6	6	0,6	0,6	1	0,75
ARTIGO15	10	4	3	0,2727	0,3	0,75	0,4286
ARTIGO16	10	11	9	0,75	0,9	0,8182	0,8571
ARTIGO17	10	6	6	0,6	0,6	1	0,75
ARTIGO18	9	3	3	0,3333	0,3333	1	0,5
ARTIGO19	10	6	5	0,4545	0,5	0,8333	0,625
ARTIGO20	10	6	6	0,6	0,6	1	0,75
ARTIGO21	10	6	6	0,6	0,6	1	0,75
ARTIGO22	9	5	5	0,5556	0,5556	1	0,7143
ARTIGO23	9	5	5	0,5556	0,5556	1	0,7143
ARTIGO24	10	5	5	0,5	0,5	1	0,6667
ARTIGO25	10	5	5	0,5	0,5	1	0,6667
ARTIGO26	10	3	3	0,3	0,3	1	0,4615
ARTIGO27	10	4	4	0,4	0,4	1	0,5714
ARTIGO28	10	5	5	0,5	0,5	1	0,6667
ARTIGO29	10	5	5	0,5	0,5	1	0,6667
ARTIGO30	6	4	3	0,4286	0,5	0,75	0,6
ARTIGO31	5	3	3	0,6	0,6	1	0,75
ARTIGO32	9	6	3	0,25	0,3333	0,5	0,4

ARTIGO33	10	5	3	0,25	0,3	0,6	0,4
ARTIGO34	10	5	3	0,25	0,3	0,6	0,4
ARTIGO35	5	5	4	0,6667	0,8	0,8	0,8
ARTIGO36	10	5	5	0,5	0,5	1	0,6667
ARTIGO37	10	8	8	0,8	0,8	1	0,8889
ARTIGO38	7	5	4	0,5	0,5714	0,8	0,6667
ARTIGO39	10	5	4	0,3636	0,4	0,8	0,5333
ARTIGO40	9	6	4	0,3636	0,4444	0,6667	0,5333
ARTIGO41	10	6	6	0,6	0,6	1	0,75
ARTIGO42	10	5	4	0,3636	0,4	0,8	0,5333
ARTIGO43	10	5	5	0,5	0,5	1	0,6667
ARTIGO44	10	4	4	0,4	0,4	1	0,5714
ARTIGO45	10	10	6	0,4286	0,6	0,6	0,6
ARTIGO46	10	6	4	0,3333	0,4	0,6667	0,5
ARTIGO47	9	6	5	0,5	0,5556	0,8333	0,6667
ARTIGO48	10	8	6	0,5	0,6	0,75	0,6667
ARTIGO49	10	5	5	0,5	0,5	1	0,6667
ARTIGO50	10	5	5	0,5	0,5	1	0,6667
ARTIGO51	7	5	3	0,3333	0,4286	0,6	0,5

ARTIGO52	10	6	6	0,6	0,6	1	0,75
ARTIGO53	10	7	5	0,4167	0,5	0,7143	0,5882
ARTIGO54	8	7	5	0,5	0,625	0,7143	0,6667
ARTIGO55	10	5	4	0,3636	0,4	0,8	0,5333
ARTIGO56	10	7	7	0,7	0,7	1	0,8235
ARTIGO57	10	7	7	0,7	0,7	1	0,8235
ARTIGO58	10	5	4	0,3636	0,4	0,8	0,5333
ARTIGO59	10	4	3	0,2727	0,3	0,75	0,4286
ARTIGO60	6	4	4	0,6667	0,6667	1	0,8
ARTIGO61	10	7	6	0,5455	0,6	0,8571	0,7059
ARTIGO62	10	8	6	0,5	0,6	0,75	0,6667
ARTIGO63	10	6	6	0,6	0,6	1	0,75
ARTIGO64	10	7	5	0,4167	0,5	0,7143	0,5882
ARTIGO65	10	6	5	0,4545	0,5	0,8333	0,625
ARTIGO66	10	7	5	0,4167	0,5	0,7143	0,5882
ARTIGO67	10	6	6	0,6	0,6	1	0,75
ARTIGO68	10	6	3	0,2308	0,3	0,5	0,375
ARTIGO69	10	7	4	0,3077	0,4	0,5714	0,4706
ARTIGO70	10	5	4	0,3636	0,4	0,8	0,5333

ARTIGO71	10	8	8	0,8	0,8	1	0,8889
ARTIGO72	10	6	5	0,4545	0,5	0,8333	0,625
ARTIGO73	10	6	5	0,4545	0,5	0,8333	0,625
ARTIGO74	8	5	4	0,4444	0,5	0,8	0,6154
ARTIGO75	10	7	5	0,4167	0,5	0,7143	0,5882
ARTIGO76	10	7	6	0,5455	0,6	0,8571	0,7059
ARTIGO77	10	6	5	0,4545	0,5	0,8333	0,625
ARTIGO78	10	8	5	0,3846	0,5	0,625	0,5556
ARTIGO79	10	5	4	0,3636	0,4	0,8	0,5333
ARTIGO80	10	6	4	0,3333	0,4	0,6667	0,5
ARTIGO81	10	8	6	0,5	0,6	0,75	0,6667
ARTIGO82	8	5	4	0,4444	0,5	0,8	0,6154

Fonte: dados da pesquisa (2021).

Em seguida, na tabela 5, são apresentados os principais resultados da análise de IA por atribuição do *Corpus* B. Identificam-se os termos obtidos por meio da IA atribuídos pelo MAUI, que apresentam os valores descritos na tabela para o cálculo dos índices de precisão, revocação e medida F. Dessa maneira, os resultados da tabela 8 contemplam os seguintes campos: Termos Extraídos do Maui, Termos Indexados Manualmente, Número de Termos Comuns, Consistência, Precisão, Revocação e Medida F.

Baseado nesses resultados, observa-se o mínimo de três termos indexados e o máximo de 11 termos atribuídos pela Indexação Manual, enquanto o Maui atribuiu um mínimo de cinco e um máximo de 10 termos indexados automaticamente. Esses valores resultam numa média de seis descritores atribuídos por documento para a Indexação Manual e uma média de nove descritores atribuídos por documento para a Indexação do Maui.

Para facilitar o processo de compreensão com relação às métricas associadas à Tabela 7, foi elaborada a tabela 5 com os valores das médias para cada métrica identificada no *Corpus* A. Como pode-se observar a seguir:

Tabela 5 - Média das Métricas Obtidas pelo Maui no *Corpus* B

	Número de termos do Maui	**Número de termos da Indexação Manual**	**Nº de termos comuns com a Indexação Manual**	**Consistência**	**Precisão**	**Revocação**	**Medida F**
Mínimo	5	3	3	25%	30%	62%	37%
Máximo	10	11	7	80%	70%	100%	88%
Média	9	6	5	47%	51%	84%	62%
Desvio Padrão	1,3133	1,5354	1,3199	0,1327	0,1307	0,1461	0,1224

Fonte: Adaptado de Silva (2020)

Seguidamente, em relação à consistência, percebe-se uma média de 47% nos Índices de Consistência da Indexação, variando entre 25% e 80%. Já, em relação à Medida F, foi obtido um índice de 62%, esse valor foi obtido graças à média harmônica das medidas de precisão e de revocação. Ademais, percebe-se 51% no índice de precisão e 84% no de Revocação. Semelhantemente aos argumentos apresentados na análise do *corpus* A, os resultados demonstraram alinhamento aos achados de Narukawa, Gil Leiva e Fujita (2009) e Bandim e Correa (2017), sendo, inclusive, superiores nas métricas obtidas.

O processo de avaliação dos resultados obtidos só se fazem necessários, se os resultados esperados não contemplarem os índices mínimos necessários para uma eficiente IA. Tais elementos devem ser levados em consideração no processo de análise das métricas obtidas. Caso os resultados não se enquadrassem no padrão estabelecido, os processos de organização da informação e do conhecimento deveriam ser reaplicados visando melhor padronizar os dados selecionados, inclusive, revisando o papel dos tesauros e das ferramentas de processamento de linguagem natural, com o propósito de reduzir a dispersão do *corpus* e torná-lo adequado aos índices mínimos de precisão, revocação, consistência e medida F.

É importante destacar também que a descrição do conjunto de termos atribuídos pelo sistema de indexação e indexados manualmente, foi feita em colunas distintas e está presente nos Apêndices D e E desta tese. Esse processo de descrição é um procedimento que deve ser realizado por um especialista da área de conhecimento pertencente aos documentos indexados e à linguagem controlada escolhida. O especialista deverá descrever cada conjunto de termo, para cada artigo, e iniciar um processo de análise cuidadosa, observando aspectos de consistência textual, levando em consideração cada elemento do processo e os fatores intervenientes que podem afetar a atribuição de cada termo escolhido. A lista de termos indexados automaticamente para cada artigo do *corpus* A e *corpus* B, utilizados nesta pesquisa, pode ser visualizado no Apêndice C desta pesquisa.

Logo, esse procedimento é relevante, pois proporciona a devida comparação de atribuição dos termos indexados, verificando os possíveis erros com relação à IM de cada artigo, além do processo de comparação de todos os termos obtidos ao final da IA. Assim, os termos logram ser comparados analisando as possíveis inconsistências, procurando verificar as possíveis causas que levam às divergências entre os diferentes procedimentos de Indexação. A lista de termos comparados para cada artigo do *corpus* A e *corpus* B, utilizados nesta pesquisa, pode ser visualizado nos Apêndices D e E deste documento.

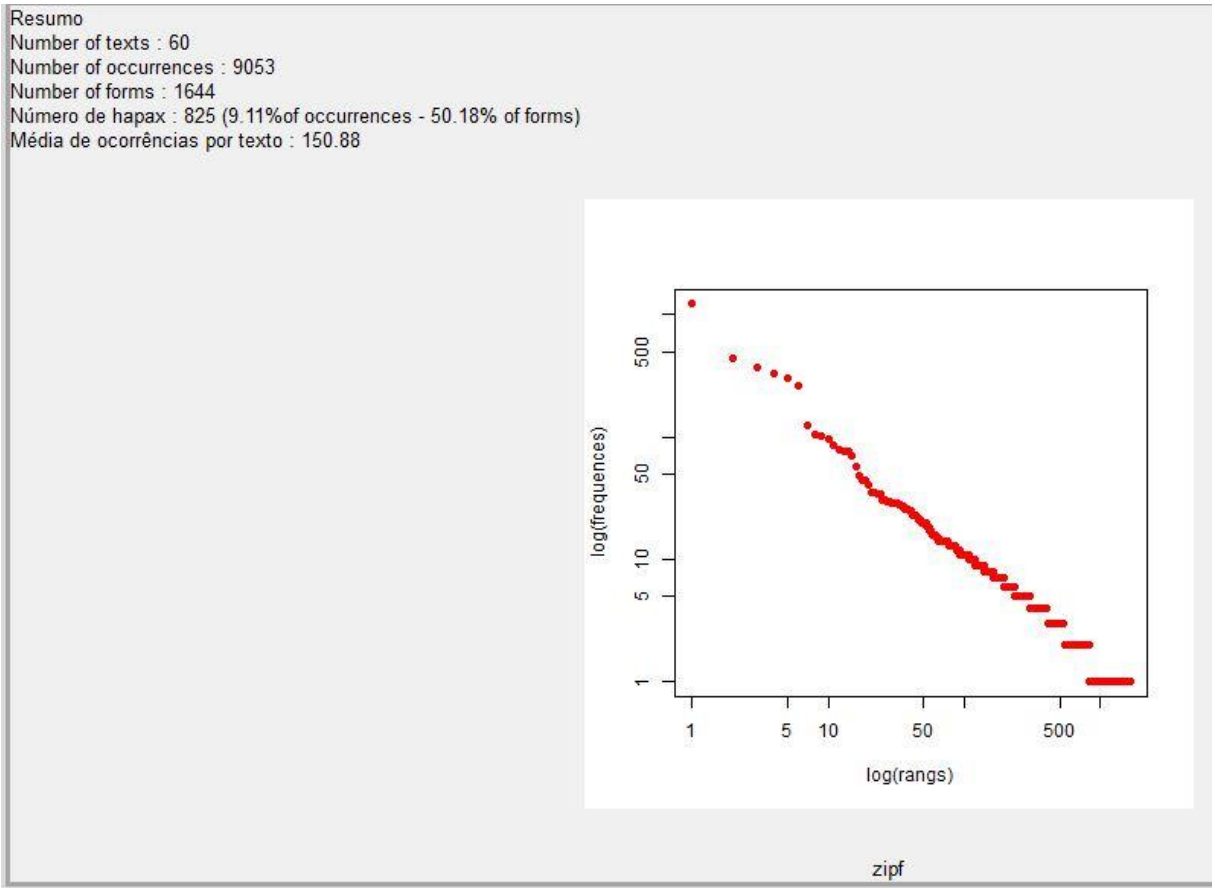
Portanto, ao serem analisadas as métricas da IA, envolvidas nesta última etapa do processo de IA, os especialistas podem verificar os principais erros atribuídos em cada processo de atribuição. Comparando os acertos e os erros. Além disso, são capazes de levantar as possíveis causas que tornam o processo de IA complexo, levando em consideração que estes procedimentos podem sofrer influências específicas devido a fatores não identificados. Portanto, com base nos parâmetros de qualidade da Indexação, levantados pelos Indicadores de Consistência, Revocação, Precisão e Medida F, será possível que o especialista possa analisar, de forma criteriosa, cada etapa envolvida no processo.

Posteriormente, com a aplicação do Modelo de EMI adotado para a elaboração dos IT, foi possível visualizar os termos mais representativos, semanticamente, dentro do contexto analisado. Esta parte da etapa ocorre nas seções 4.2 e 4.3 a seguir.

4.2 Análises Estatísticas dos Corpora

A primeira proposta de análise é um resultado estatístico obtido por meio do *software* IRAMUTEQ. Esse tipo de análise procura demonstrar a aplicação da Lei de Zipf (ZIPF, 1949), observando o grau de dispersão das palavras em ambos os *corpora*. Para sua elaboração foram levados em consideração os parâmetros mínimos para a identificação do documento e das estruturas textuais existentes no *corpus*. Seguiu-se assim, a metodologia de análise proposta por Ratinaud e Marchand (2009), o que resultou no gráfico de classificação e frequência, em escala logarítmica, no gráfico 6, a seguir.

Gráfico 6 - Estatística do *corpus* A com os Metadados (Título, Resumo e Palavras-Chave do Autor)



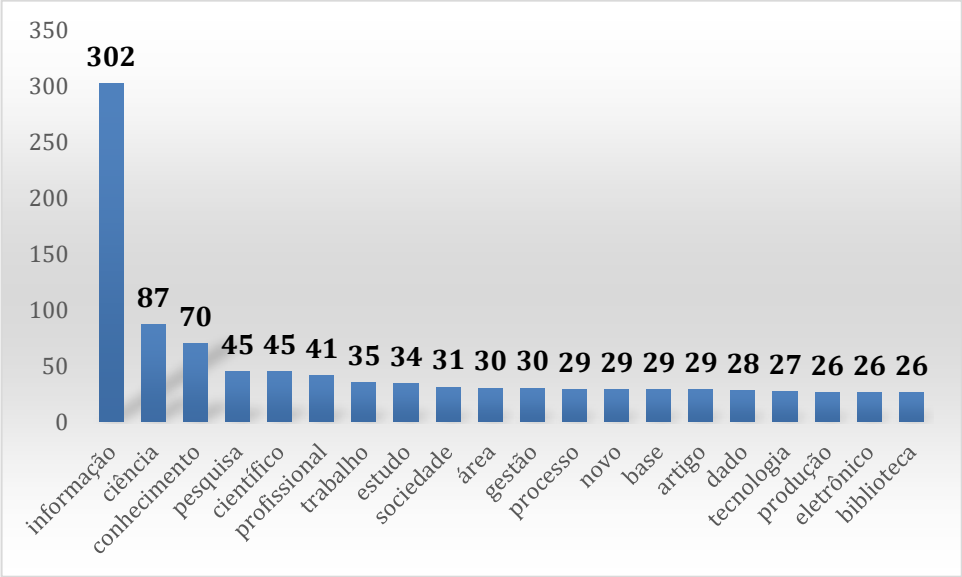
Fonte: dados da pesquisa (2021).

A primeira análise realizada é a aplicação da Lei de Zipf no *corpus* A. Essa lei busca, de acordo com Urbizagástegui Alvarado (1984), calcular a frequência das palavras. Além

disso, Araújo (2006) pondera sobre a lei, discriminando a relação existente entre palavras num determinado contexto. Assim, é verificada uma correlação entre o número de palavras diferentes e a frequência de seu uso, existindo uma regularidade na seleção e uso das palavras onde um pequeno número de palavras é utilizado mais frequentemente e um grande número tem menor utilização. Mais especificamente, na análise acima, é identificado que existem 60 documentos contendo 9053 ocorrências de termos para 1644 tipos de palavras. Entretanto, cerca de 9,11% das ocorrências (825 palavras), aparecem apenas uma vez nos documentos. Em relação aos tipos de palavras, isto representa 50,18%. Sendo assim, pode-se afirmar que, pelo menos metade dos tipos de palavras aparece apenas uma vez, sugerindo uma configuração de dispersão, tendo em vista que a outra metade responde por 89,89% das vezes.

Essa relação é comprovada pelo que foi afirmado por Araújo (2006), quando sugere uma utilização regular na seleção e no uso de certas palavras, sendo utilizadas, sobremaneira, mais frequentemente. Ressalta-se, que para esta análise, foram consideradas as palavras unitárias, ou seja, não compostas. Os termos compostos serão abordados na etapa 5 de validação do fluxograma. Portanto, percebe-se a tendência de ocorrer um grande índice de repetição de palavras, o que pode ser comprovado no gráfico de frequência absoluta no gráfico 7, a seguir.

Gráfico 7 - 20 Palavras mais frequentes do *corpus* A



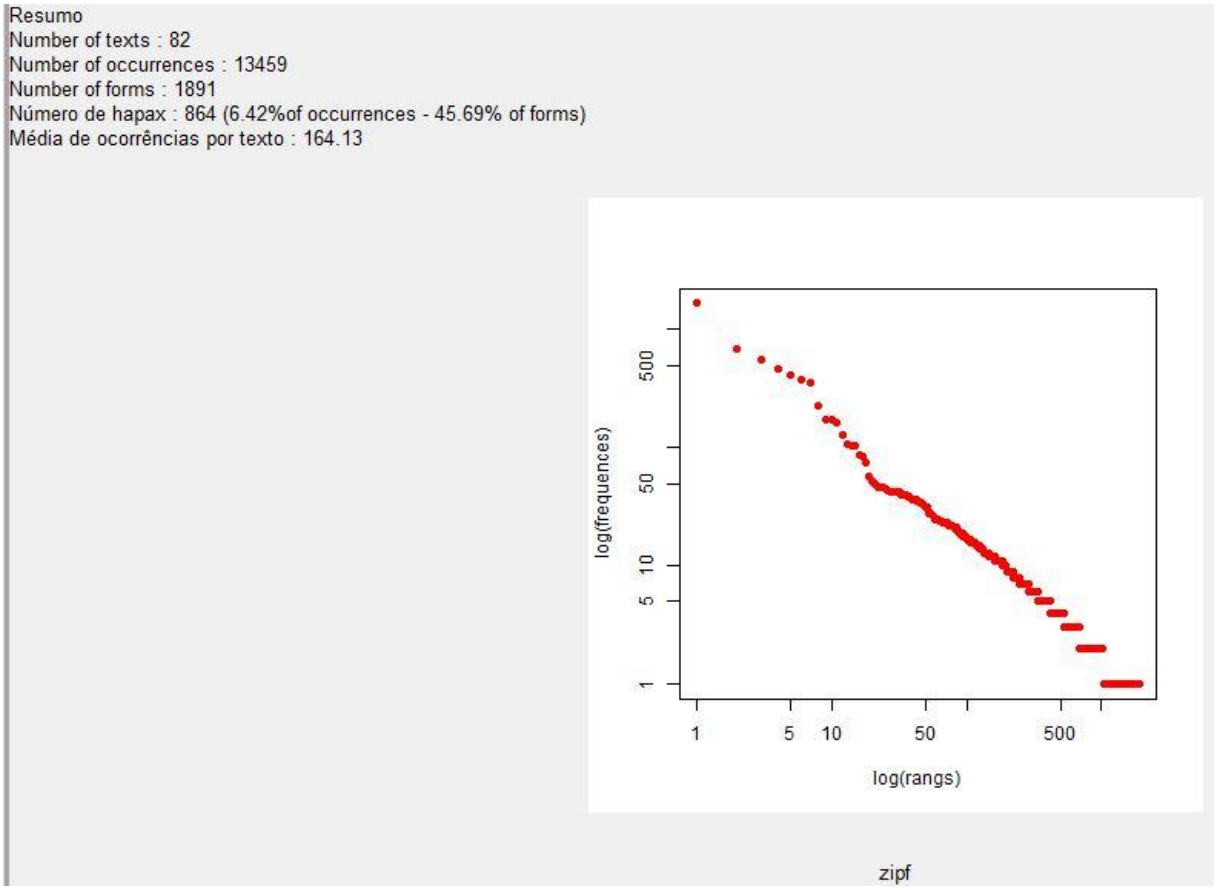
Fonte: dados da pesquisa (2021).

O gráfico 7, de frequência de palavras do *corpus* A, comprova a aplicação da lei de Zipf quando identifica um pequeno conjunto de palavras com alta ocorrência no *corpus* e um grande número com ínfima frequência. Identificaram-se as palavras “informação”, “ciência” e “conhecimento” como os termos dominantes, evidenciando o núcleo temático do *corpus*

estudado. A partir da aplicação da lei de Zipf, evidenciou-se que poucas palavras conseguem representar, semanticamente, um documento ou um conjunto de documentos, como foi afirmado por Araújo e Urbizagástegui Alvarado.

Para permitir a continuidade das análises propostas, foi elaborado o gráfico 8, estatísticas do *corpus* B, para a comprovação dos fenômenos identificados no *corpus* A.

Gráfico 8 - Análise Estatística do *corpus* B

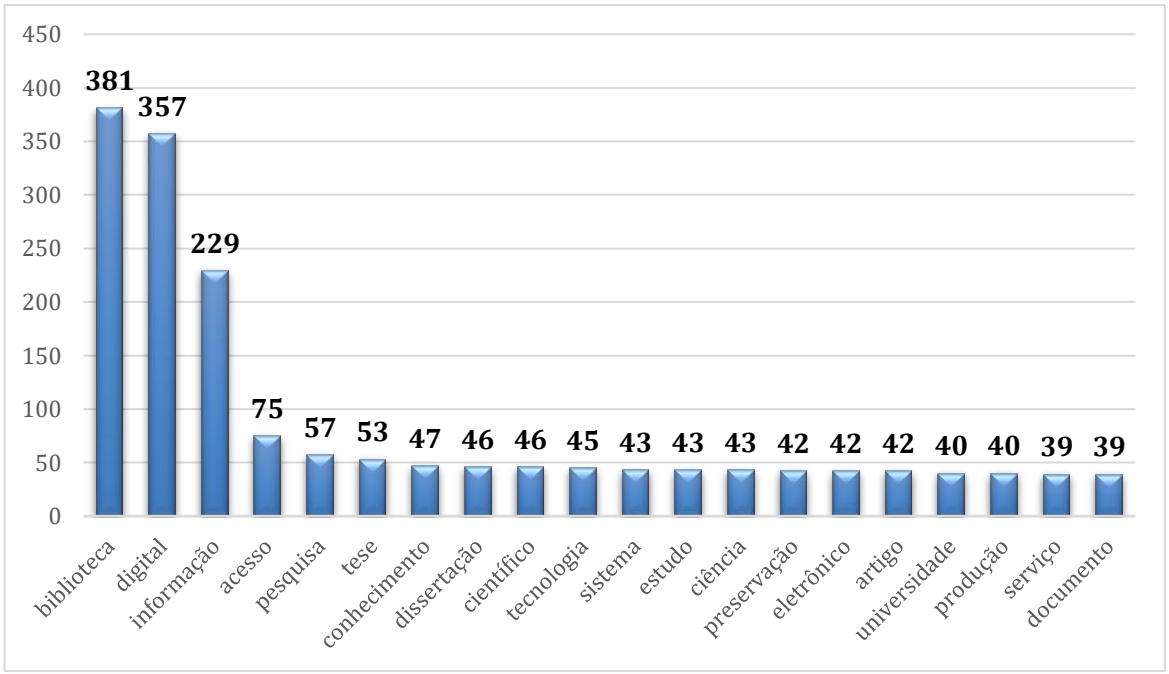


Fonte: dados da pesquisa (2021).

De maneira semelhante, como afirmou Araújo (2006), o gráfico de análise Estatística do *corpus* B apresenta correlação entre as frequências das palavras num determinado contexto. Neste sentido, verifica-se uma conexão entre a ocorrência de palavras díspares e a frequência na sua utilização. Existe portanto, uma regularidade na escolha de certos termos onde um baixo número de vocábulos é utilizado mais frequentemente. Na análise acima é identificado que existem 82 documentos analisados, contendo 13.459 ocorrências de termos para 1891 tipos de palavras. Entretanto, cerca de 6,42% das ocorrências, ou 864 palavras, aparecem apenas uma vez nos documentos, para 45,69% das formas, ou seja, 45% das palavras utilizadas são citadas uma única vez, enquanto a outra metade é citada 93% das vezes. Portanto verifica-se a mesma tendência do *corpus* A de ocorrer uma grande incidência

na repetição de palavras, o que pode ser comprovado no gráfico 9 de frequência absoluta a seguir.

Gráfico 9 - 20 Palavras mais frequentes do *corpus* B



Fonte: dados da pesquisa (2021).

Como ocorre no *corpus* A, o gráfico 9, ratifica a aplicação da lei de Zipf quando identifica um pequeno conjunto de palavras com alta ocorrência no *corpus*. Percebe-se as palavras "biblioteca", "digital" e "informação" como os vocábulos mais frequentes, comprovando a relação de dependência semântica do *corpus* analisado a essas palavras supracitadas. Portanto, mais uma vez comprova-se que uma pequena quantidade de termos tem a capacidade de representar o conteúdo conceitual de um determinado conjunto de documentos.

4.3 Análise das Etapas do Estudo Métrico

Etapa 1 do fluxograma - Formatar os *corpora* A e B e adaptá-los para o formato exigido no Iramuteq

Nesta etapa do processo de elaboração dos *corpora* foi necessário formatar os arquivos para garantir que eles possuíssem o formato exigido para serem lidos pelo *software*. Os segmentos de texto (Título, Resumo e Palavras-chave) de cada artigo foram copiados e inseridos num arquivo de bloco de notas, para serem salvos no formato de texto sem formatação e com codificação UTF-8 sem "BOM". Em seguida, foi atribuída uma

nomenclatura textual (título) para cada Artigo no seguinte formato: **** *Artigo01. Nesse aspecto, foi possível identificar cada segmento textual como parte do referido artigo utilizado na análise textual. Para o *corpus* A foram atribuídas 60 nomenclaturas para os 60 artigos, e para o *corpus* B foram atribuídas 82 nomenclaturas. Após a inserção dos metadados foi feita uma verificação visual de todos os segmentos textuais em cada *corpus*, verificando se não haveria faltado nenhum metadado atribuído ou nomenclatura textual inserida ou possíveis erros identificáveis. Em seguida, após esse processo foi implementada a análise estatística.

Etapa 2 do fluxograma - Realizar análise estatística em cada *corpus*

Em seguida, realizou-se uma análise estatística para demonstrar a aplicação da Lei de Zipf, observando o grau de dispersão das palavras em ambos os *corpora*. Nesse âmbito, foi possível comparar a frequência e a ocorrência das palavras, pois o *software* fornece o número de textos e segmentos de textos, ocorrências, frequência média das palavras, bem como a frequência total de cada forma.

Etapa 3 do fluxograma - Definir as configurações e parâmetros necessários para a elaboração da análise de similitude

Nesta fase é importante estabelecer as configurações necessárias para promover a análise correta da coocorrência de palavras. Para esta tese foi adotado o método de visualização do algoritmo de Kamada e Kawai, que fornece uma modelagem dos vértices por meio de um sistema de visualização que prioriza o tamanho proporcional à distância entre os vértices, nesse contexto, é possível isolar melhor os vértices e formar arestas com tamanhos similares e com pouco cruzamento entre elas (LIMA, 2017).

Nesse cenário, Kamada e Kawai propuseram em 1989 um método gráfico que tem como propósito realizar representações gráficas de estruturas computacionais complexas, dentre as quais, as redes de Petri e os diagramas de estado. Para elaborar essas visualizações, Kamada e Kawai sugerem um algoritmo que é baseado no método das forças. Esse método adota os parâmetros desejados, levando em consideração as distâncias mais longas entre os vértices, simulando as forças por meio de um efeito de mola. Assim, as distâncias euclidianas entre os nós são consideradas posicionando os nós de forma a não sobrepor os vértices e arestas com frequência (MARTÍNEZ, 2014). Após os ajustes e configurações necessárias, seguiu-se para a etapa seguinte da análise.

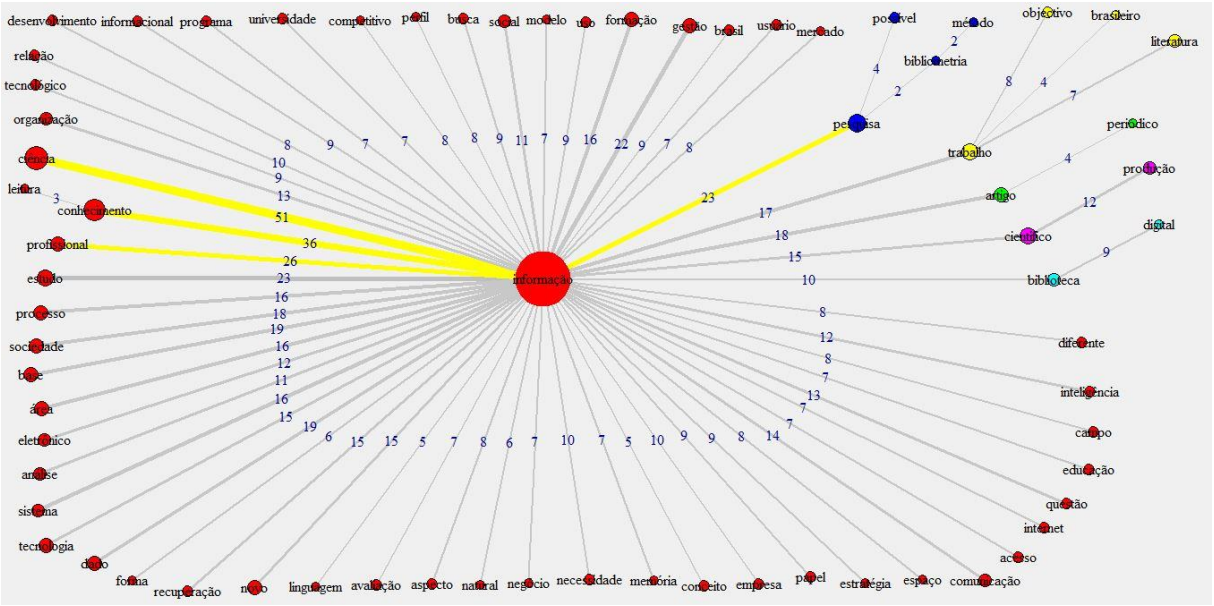
Etapas 4 e 5 do fluxograma - Realizar análise de similitude por visualização em rede e gerar os gráficos, e, realizar análise de cada gráfico obtido, inserindo as descrições e argumentos necessários.

Essas duas etapas foram importantes, pois possibilitaram a elaboração dos parâmetros necessários para construção do gráfico de análise de similitude e sua respectiva análise. De acordo com Marchand e Ratinaud (2012), essa análise baseia-se na teoria dos grafos[39] que é frequentemente adotada por pesquisadores que trabalham com interações sociais. Ela possibilita a identificação das coocorrências entre as palavras, trazendo indicações do grau de conexidade entre os termos, auxiliando na identificação da estrutura do conteúdo do *corpus* textual. A consolidação desta etapa ocorre após a devida configuração e elaboração dos gráficos, obtendo a visualização necessária para realização das análises. Os gráficos de análise de similitude foram construídos com o propósito de observar o comportamento temático das palavras dentro de cada *corpus*. Esse tipo de gráfico é interessante, pois apresenta possíveis proximidades semânticas entre as palavras escolhidas para cada *corpus*, e, ressalta elementos quantitativos com base na ocorrência das relações e dos atores, sendo isto, expressado pelo tamanho dos nós e dos vínculos.

A figura 12, a seguir, foi elaborada com os 70 termos mais frequentes e relevantes do *corpus* A, de uma ocorrência que vai de 11 a 302 de cada palavra, da menos frequente para a mais frequente.

[39] De acordo com FEOFILOFF, KOHAYAKAWA e WAKABAYASH (2011, p. 8) um grafo é um par (V, A) em que V é um conjunto arbitrário e A é um subconjunto de V. Os elementos de V são chamados vértices e os de A são chamadas arestas. Uma aresta como {v, w} é denotada simplesmente por vw ou por wv . A aresta vw incide em v e em w e que v e w são as pontas da aresta. Se vw é uma aresta, afirma-se que os vértices v e w são vizinhos ou adjacentes. De acordo com essa definição, um grafo não pode ter duas arestas diferentes com o mesmo par de pontas (ou seja, não pode ter arestas "paralelas"). Também não pode ter uma aresta com pontas coincidentes (ou seja, não pode ter "laços").

Figura 12 - Análise de Agrupamentos (Clusters) do *corpus* A com os Metadados (Título, Resumo e Palavras-Chave do Autor)



Fonte: dados da pesquisa (2021).

Nesse tipo de visualização, é possível identificar um polo central (com o termo “Informação”) que possui em seu entorno diversos termos tematicamente conectados. Na análise quantitativa foi identificado que o termo “Informação” é a palavra mais frequente, apresentando 302 ocorrências. Observa-se, a partir deste ponto de vista, que os termos se ligam a outros por meio de fios cinza-claro. Na teoria dos grafos, esses fios são denominados arestas. A espessura dessas arestas representa o grau de conexão entre as palavras observadas. Nesse caso, observa-se que “Informação” se conecta, principalmente, com os termos: “Ciência”, “Conhecimento” “Profissional” e “Pesquisa” (vide Figura 12).

Nota-se que algumas palavras aparecem em agrupamentos externos ao agrupamento central em vermelho. Em relação à “Informação”, percebe-se que está mais intensamente conectada à palavra “ciência”, devido a maior espessura da aresta. Isso ocorre graças à conjuntura temática do *corpus* estudado, os 60 artigos analisados neste *corpus* pertencem ao campo da “Ciência da Informação”, e dentre os elementos desse campo são estudadas áreas fenomenológicas como a própria Informação e o próprio conhecimento. Por esse motivo, a palavra “conhecimento” conecta-se com “informação” apresentando um índice de força de ocorrência no valor de 36.

Outra análise importante foi a clusterização das palavras, visando identificar os termos que se agrupam por critérios relacionais e de similitude no grafo. Com isto, perceberam-se agrupamentos compostos por palavras distintas, a saber: 1- O primeiro agrupamento, na cor azul escura, possui 4 palavras identificadas nos círculos que se conectam entre si (“Pesquisa”

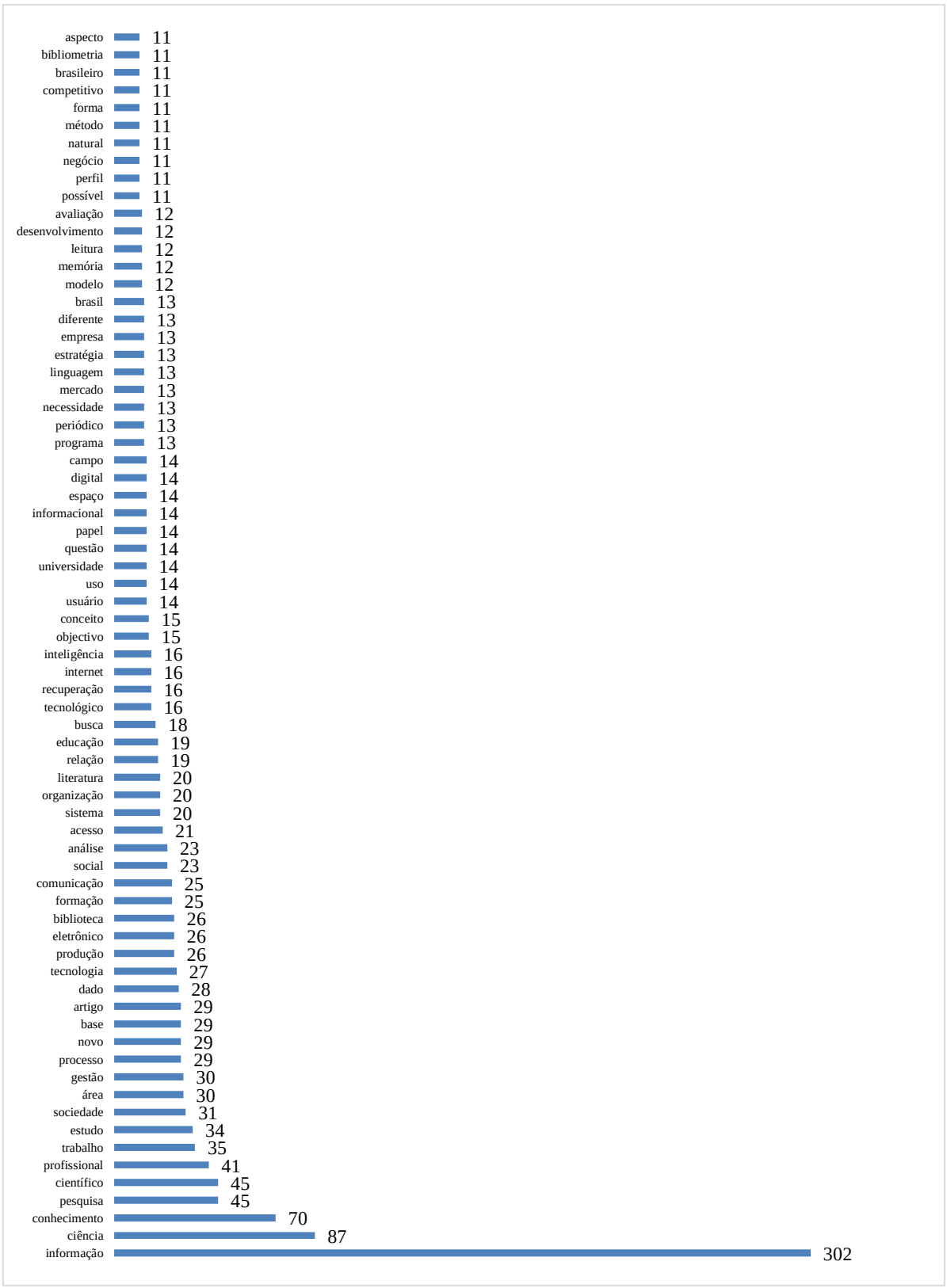
com 45 ocorrências, “bibliometria” com 11 ocorrências, “possível” com 11 ocorrências e “método” com 11 ocorrências). Essas palavras ocorrem, com maior frequência, próximas à palavra “Informação”. Todas são representativas, em maior ou menor grau, porém, destaca-se a palavra pesquisa conectando o grupo ao termo “informação”. Essa relação deve ocorrer pois muitos artigos são voltados para área de Pesquisa em Ciência da Informação, bibliometria e descrevem os métodos de pesquisa adotados nos estudos. 2- O segundo agrupamento, está representado na cor amarela com 4 palavras conectadas (trabalho, 35 ocorrências, objetivo, 15 ocorrências, brasileiro, 11 ocorrências e literatura com 20 ocorrências), essas palavras se relacionam de forma mais próxima entre elas, porém, quase de forma simultânea, se conectam ao termo Informação.

Essa correlação pode ser justificada pela descrição dos procedimentos metodológicos dos estudos, pois, as locuções “Objetivo do trabalho” e “Literatura brasileira” são comumente utilizados para descrever trechos dos procedimentos descritos, portanto, entende-se que a relação existente é justificável no que diz respeito à representação visual vigente. 3- O terceiro agrupamento observado, em verde, apresenta dois termos que se relacionam de forma mais intensa, tanto no aspecto de frequência quanto no de coocorrência. (Artigo com 29 ocorrências e Periódico com 13 ocorrências) se apresentam como palavras que ocorrem de maneira muito frequente na descrição dos objetos utilizados nas análises. Percebe-se que esses termos ocorrem frequentemente pois as pesquisas em CI, envolvem, em muitos momentos, os artigos de periódicos como objeto principal das análises. 4- Posteriormente, o agrupamento em rosa, apresenta 2 termos (Científico com 26 ocorrências e Produção com 45 ocorrências). Nesse agrupamento, os termos surgem como palavras mais coocorrentes.

Esses fatos podem ser justificados pela proximidade entre as palavras e sua relação com a área acadêmica no âmbito da produção científica em CI. Além disso, essas palavras se relacionam diretamente com a área temática relacionada aos EMI, aplicados na metodologia desta pesquisa. 5- Por fim, observa-se o agrupamento representado pelas palavras (biblioteca com 26 ocorrências e digital com 14 ocorrências), na cor azul clara, essas palavras conectam-se fortemente pois são comumente utilizadas como locução nominal “biblioteca digital”, também conhecida como “biblioteca eletrônica”, “biblioteca virtual”, relacionadas à criação, aquisição, distribuição e armazenamento de documentos digitais. Esse domínio pode ser facilmente encontrado na constelação temática que compõe à área de CI.

Como forma de facilitar a identificação dos índices de ocorrência das palavras no *corpus* A, foi elaborado o gráfico 10 a seguir, objetivando expressar os termos mais representativos presentes no gráfico de similitude.

Gráfico 10 – Palavras mais frequentes do *corpus* A (Indexação do Autor)

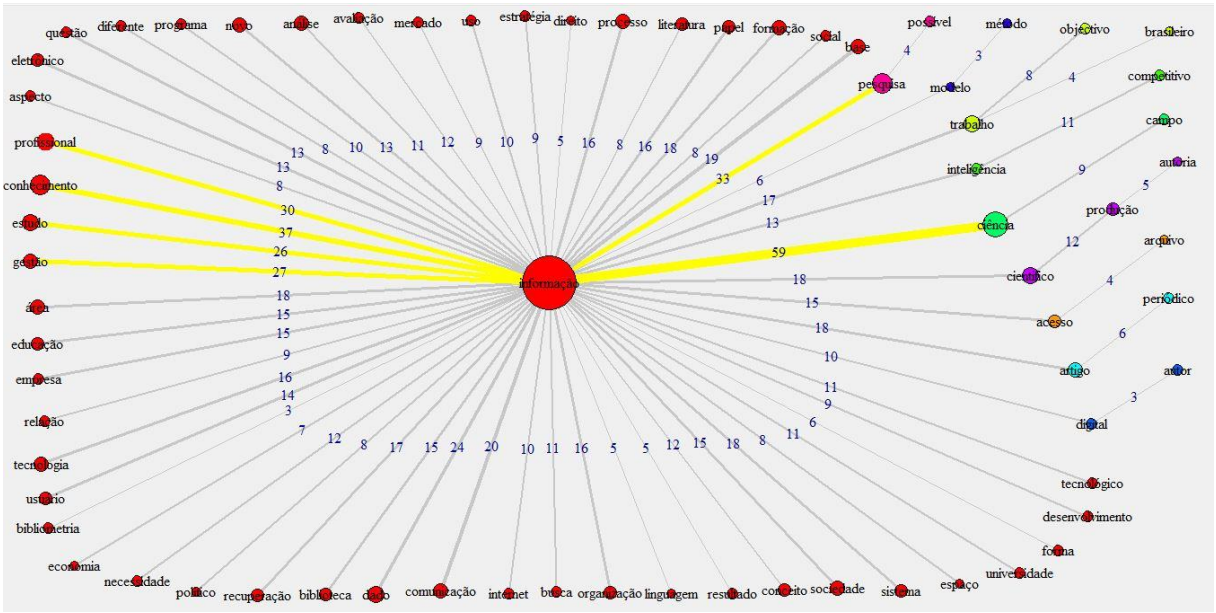


Fonte: dados da pesquisa (2021).

Em seguida, optou-se por elaborar um segundo gráfico de similitude (Figura 13), com o objetivo de analisar o comportamento dos termos e suas relações temáticas substituindo os termos indexados manualmente pelos termos indexados pelo Maui. Essa análise tem o principal propósito de identificar essa variação na correlação entre os termos, além da variação em termos quantitativos no *corpus*.

A figura 13 foi estruturada com os 70 termos mais frequentes e relevantes do *corpus*, de uma ocorrência que vai de 10 a 337 de cada palavra, da menos frequente para a mais frequente.

Figura 13 -Análise de Agrupamentos (Clusters) do *corpus* A com os Metadados (Título, Resumo e Palavras-Chave do Maui)



Fonte: dados da pesquisa (2021).

Na figura 13, é possível identificar 1 polo central (com o termo “Informação”) que possui relação com diversos termos (na cor vermelha), conectando-se tematicamente. Na análise de frequência foi identificado que o termo “Informação” é a palavra mais frequente, apresentando 337 ocorrências. Observa-se, a partir deste ponto de vista, que os termos se ligam a outros por meio de fios cinza-claro, como foi citado anteriormente, para a teoria dos grafos esses fios são denominados arestas. A espessura dessas arestas representa o grau de conexão entre as palavras observadas. Nesse contexto, observa-se que “Informação” se conecta com termos como: “Ciência”, “Conhecimento”, “Pesquisa”, “Profissional” e “Gestão” (vide Figura 13).

Ao comparar os resultados do Maui (IA) com as palavras dos autores, nota-se que a principal diferença é a inserção do termo “Gestão” vinculado à “Informação”, advindo dos processos de IA. A Gestão da Informação consiste em identificar e potencializar os recursos

informacionais de uma organização, ensinando-a a aprender e adaptar-se às mudanças ambientais (TARAPANOFF, 2001). No mais, nota-se que os resultados são bastante parecidos, sendo predominantes as discussões sobre Ciência, Pesquisa em Ciência da Informação e Conhecimento, numa perspectiva mais acadêmica.

Nota-se que algumas palavras aparecem em agrupamentos externos ao cluster central em vermelho. Em relação à "Informação", percebe-se que está mais intensamente conectada à palavra "ciência", devido a maior espessura da aresta, e ao maior grau de conexidade "59". Isso ocorre graças à conjuntura temática do *corpus* estudado, os 60 artigos analisados neste *corpus* pertencem ao campo da "Ciência da Informação", e dentre os elementos desse campo são estudadas áreas fenomenológicas como a própria Informação e conhecimento. Por esse motivo, a palavra "conhecimento" conecta-se com "informação" apresentando um índice de força de ocorrência no valor de 37.

Outra análise importante é identificada pelo grau de proximidade dos 9 agrupamentos de palavras distintos. Diferentemente do gráfico de similitude, que utiliza palavras-chave do autor (Figura 13), este gráfico possui 4 agrupamentos a mais. Dessa forma, identificam-se os seguintes agrupamentos, de acordo com a ordem de proximidade ao termo Informação: 1- O primeiro agrupamento, na cor rosa, possui 2 palavras identificadas nos círculos que se conectam entre si. Essas palavras ocorrem com maior frequência entre elas, que são: (pesquisa, com 59 ocorrências, e possível, com 11 ocorrências). Essas palavras são representativas e devem se relacionar com "Informação" devido às "possibilidades de pesquisas" para área de CI. 2- O segundo agrupamento está representado na cor azul escura com duas palavras conectadas (modelo e método, ambas com 11 ocorrências), essas palavras se relacionam de forma mais próxima entre elas, porém, quase de forma simultânea, se conectam ao termo Informação.

Essa correlação pode ser justificada descrição dos procedimentos metodológicos dos estudos, pois, as locuções "Método da CI" ou "Modelo de pesquisa da CI", ou "O modelo do método adotado" são comumente utilizados para descrever trechos dos procedimentos descritos nos artigos analisados, portanto, entende-se que a relação existente é justificável no que diz respeito à representação visual vigente. 3- O terceiro agrupamento identificado é o das palavras identificadas na cor verde claro, são elas: inteligência, com 17 ocorrências, e competitivo, com 13 ocorrências. Esses termos aproximam-se graças ao grau de conexidade entre eles, e, de forma simultânea, esse grupo de palavras tende a ser descrito se relacionando com o vocábulo "Informação". Essa relação ocorre pois o campo da "Inteligência Competitiva" é comumente utilizado como campo de estudo e análise na área da CI. 4- O

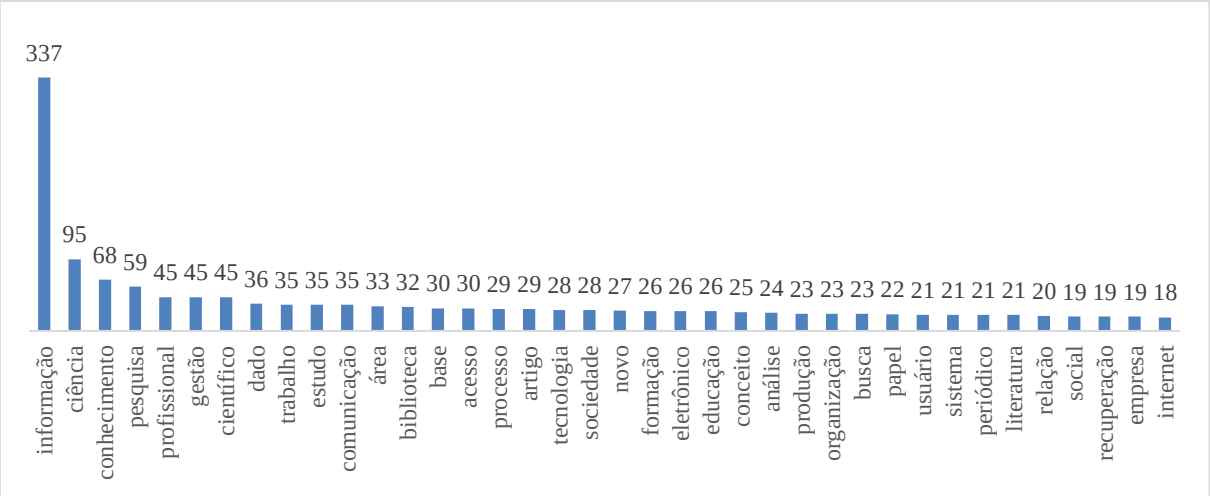
quarto agrupamento observado, em verde, apresenta 2 termos que se relacionam de forma mais intensa, tanto no aspecto de frequência quanto no de coocorrência. (Artigo com 29 ocorrências e Periódico com 13 ocorrências) se apresentam como palavras que ocorrem de maneira muito frequente na descrição dos objetos utilizados nas análises. Percebe-se que esses termos ocorrem frequentemente pois as pesquisas em Ciência da Informação, envolvem em muitos momentos, os Artigos de Periódicos como objeto principal das análises. 5- Posteriormente, identifica-se o agrupamento mais relevante, que se conecta com "Informação". Em verde claro, observa-se (Ciência, com 95 ocorrências e campo com 14 ocorrências). Esses termos são os mais relevantes, dentro da análise de similitude, pois a palavra "Ciência" apresenta um índice de conexidade, no valor de 56 em relação à palavra "Informação".

Sendo assim, muitos artigos tratam do Campo da CI como área estudo e análise, essas palavras são frequentemente citadas e justificam os valores identificados neste agrupamento. 6- Em seguida, observa-se o agrupamento representado pelas palavras (científico com 45 ocorrências, produção com 23 ocorrências e autoria com 10 ocorrências), na cor lilás, essas palavras conectam-se fortemente pois são comumente utilizadas como locução nominal "Produção Científica" e "Autoria da produção científica", esses termos compostos são áreas de estudo dos EMI e pode ser facilmente encontrada dentro do campo de pesquisa da CI. 7- Posteriormente, identificam-se 2 termos que interagem com a palavra "Informação", (Acesso com 30 ocorrências e Arquivo com 11 ocorrências).

Esses termos possuem relação temática com Informação, pois a área temática em que atuam, principalmente o "Acesso à Informação" e o "Acesso aos arquivos" são áreas importantes de estudo da CI. 8- Logo após, observa-se os termos (Artigo com 29 ocorrências e Acesso com 30 ocorrências), no agrupamento em azul claro. Essas 2 palavras possuem relevância temática pois se relacionam com "Acesso à Informação" e "Acesso à artigos e documentos". 8- Logo após, percebe-se os termos (Artigo, com 29 ocorrências e periódico com 21 ocorrências), essas palavras ocorrem muito frequentemente próximas pois indicam, possivelmente, o objeto de análise dos estudos ou o local no qual são aplicadas as metodologias desenvolvidas, em "Artigos de Periódico". 9- E por fim, ressalta-se o último agrupamento de palavras, com os termos (Digital, com 16 ocorrências e Autor, com 11 ocorrências). Esses termos, seguindo a mesma lógica, conectam-se de forma intensa, entre si, ocorrendo com certa frequência em orações devido à sua forte conexão e relação com relação à autoria dos artigos e a menção aos respectivos autores.

Com o propósito de facilitar a compreensão e a identificação dos termos mais frequentes na indexação automática do Maui, observa-se o gráfico 11. Percebe-se que "informação" continua sendo a palavra mais relevante, inclusive com um relativo aumento na sua importância, por ter sido indexada 337 vezes, ante as 302 vezes da indexação do autor.

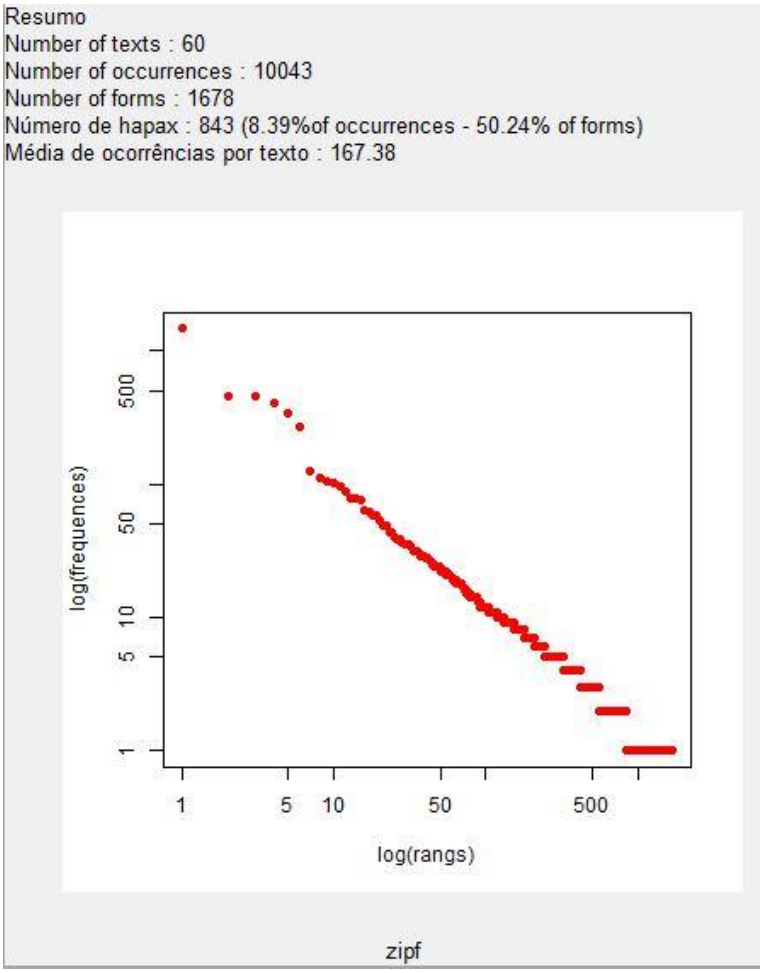
Gráfico 11 – Palavras mais frequentes do *corpus* A (Indexação do Maui)



Fonte: dados da pesquisa (2021).

Em seguida, foram elaboradas análises com bases nos descritores identificados pelos indexadores manuais. Essas análises contemplam o grau de ocorrência e frequência dos termos observados, as frequências absolutas das palavras isoladas e das palavras-chave compostas/isoladas, além de apresentar um gráfico de similitude, contemplando os principais indicadores temáticos observados. Portanto, verificam-se as análises a seguir.

Gráfico 12 - Estatística dos Indexadores Manuais – *corpus* A



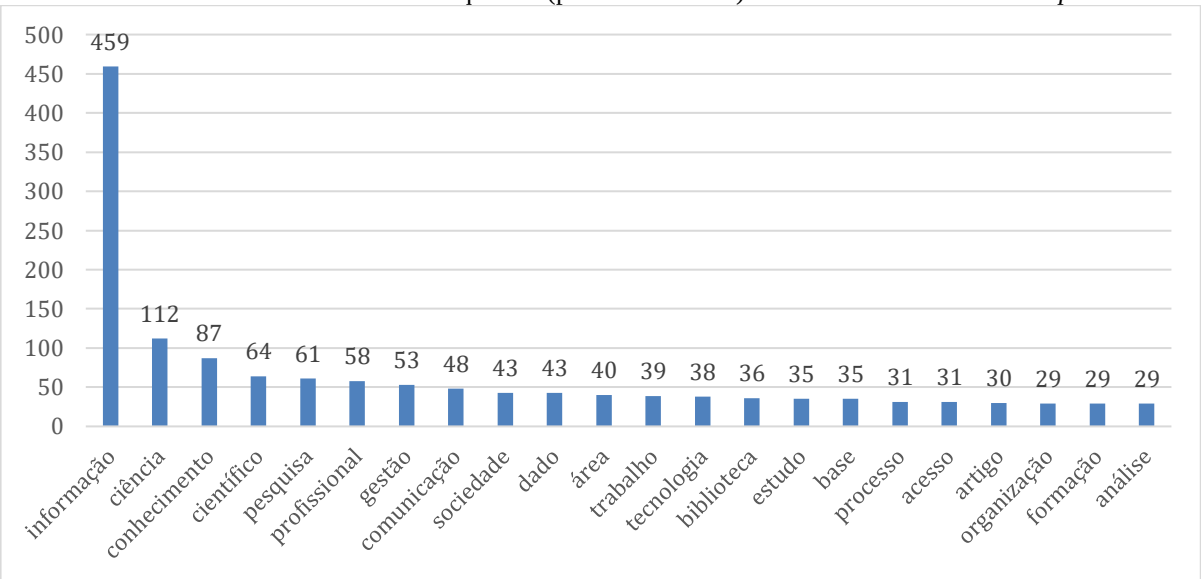
Fonte: dados da pesquisa (2021).

De forma análoga às análises anteriores, ao gráfico 12 que contempla o gráfico de análise Estatística com os termos registrados pelos indexadores manuais, do c*orpus* A, expõe as frequências e coocorrências das palavras e dos metadados presentes no *corpus*. Nesta lógica, identifica-se uma constância na predileção de determinadas palavras, no qual um pequeno número de expressões é utilizado repetidamente. Neste sentido foram considerados 60 documentos, englobando 10.043 ocorrências, para 1678 tipos de vocábulos. Entretanto, cerca de 8,39% das ocorrências, ou 843 palavras, aparecem apenas uma vez nos documentos, para 50,24% das formas, ou seja, 50% das palavras utilizadas são citadas uma única vez, enquanto a outra parte ocorre 91% das vezes. Portanto, verifica-se a mesma tendência do *corpus* A, com os descritores adotados pelos indexadores, de ocorrer uma grande incidência na repetição de palavras.

No Gráfico 13 é possível observar as palavras isoladas mais frequentes de acordo com o que foi identificado no gráfico de dispersão. Assim, é possível observar as palavras mais representativas do *corpus* A, somadas aos descritores indexados manualmente. Nessa

conjuntura, percebe-se uma relação próxima, identificável nos gráficos de frequência absoluta, anteriores, na qual surgem poucas palavras citadas frequentemente, como a palavra "informação" com 459 ocorrências, por exemplo. Assim, admite-se que este gráfico cumpre o papel de descrever os termos fora de sua estrutura relacional, visando exclusivamente expressar sua frequência absoluta no cômputo geral da pesquisa, portanto, sua leitura e análise são similares aos achados presentes no grafo, que evidenciam a predominância de "Informação", "Ciência" e "Conhecimento".

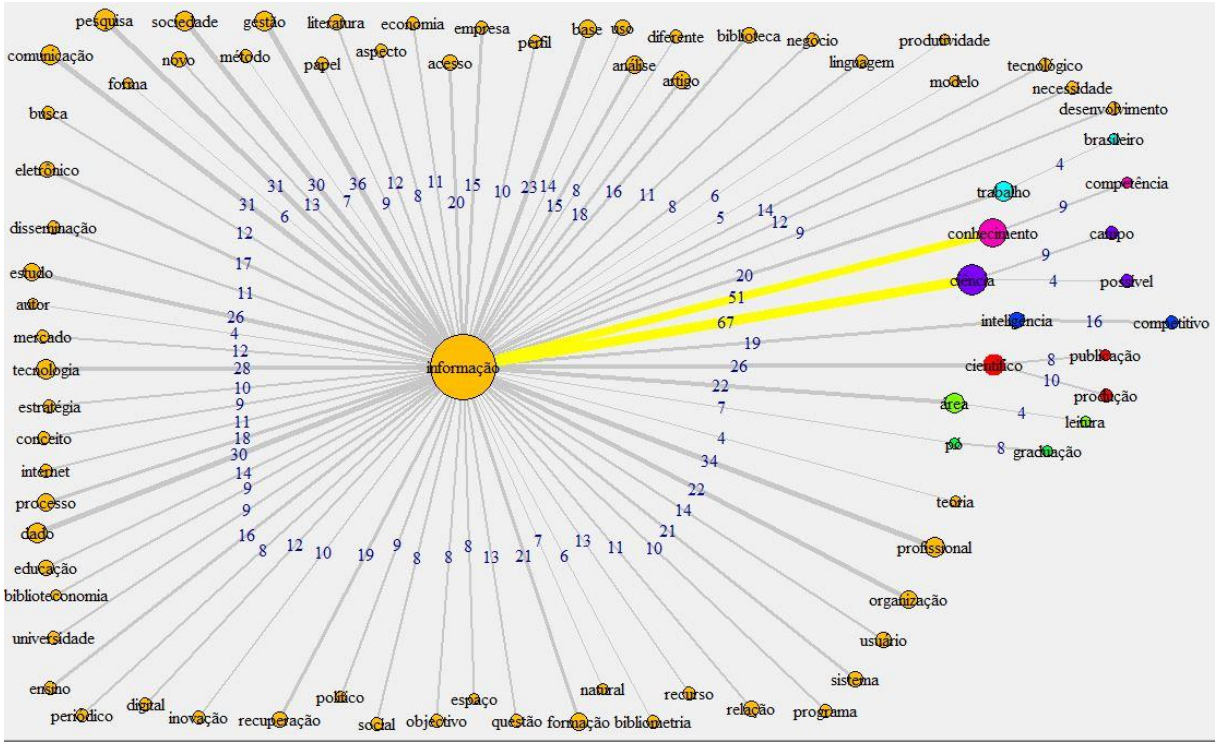
Gráfico 13 - 22 descritores mais frequentes (palavras isoladas) - Indexadores Manuais – *corpus* A



Fonte: dados da pesquisa (2021).

Posteriormente, observa-se o gráfico de similitude (figura 14). Nele é possível observar os principais agrupamentos relacionados aos indicadores temáticos dos descritores indexados pelo indexador manual no *corpus A*.

Figura 14 -Análise de Agrupamentos (Clusters) do *corpus* A com os Metadados (Título, Resumo e Palavras-Chave do Indexador Manual)



Fonte: dados da pesquisa (2021).

A figura 14, apresenta os indicadores temáticos dos principais agrupamentos e relacionamentos entre as palavras no *corpus A*, com as palavras indexadas manualmente. Para elaborar esse gráfico foram utilizadas as 80 palavras mais frequentes do *corpus* com ocorrência mínima no valor de 11 unidades. O grau de proximidade das palavras é representado pelos 8 agrupamentos de palavras distintos observados. Primeiramente, identifica-se a palavra "informação" como sendo a principal nessa análise, essa palavra se conecta com a maioria dos termos observados, inclusive por sua importância semântica no contexto textual considerado.

Nessa conjuntura, identificam-se os seguintes agrupamentos, de acordo com a ordem de proximidade ao termo Informação: 1- O primeiro agrupamento, na cor azul claro, possui 2 palavras identificadas nos círculos que se conectam entre si. Essas palavras ocorrem com maior frequência entre elas, que são: (trabalho, com 39 ocorrências, e brasileiro, com 11 ocorrências). Essas palavras são representativas e devem se relacionar com "Informação" devido às "possibilidades de pesquisas" para área de CI. 2- O segundo agrupamento, está representado na cor azul rosa com duas palavras conectadas (conhecimento, com 87 ocorrências, e competência, com 12 ocorrências), essas palavras se relacionam de forma mais próxima entre devido à proximidade temática entre "competência", "informação" e "conhecimento". A palavra conhecimento, inclusive, possui o terceiro maior nível de

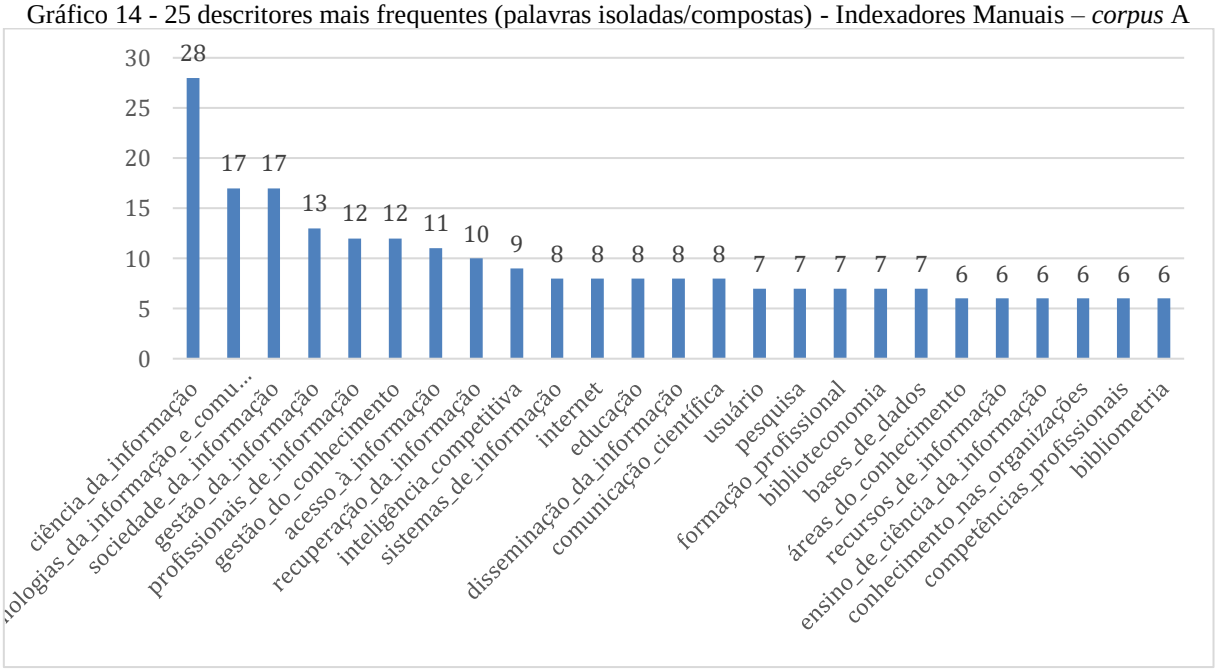
ocorrência no *corpus A*, sendo assim, o agrupamento apresenta relação temática com os termos representados. 3- O terceiro agrupamento identificado é o das palavras identificadas na lilás, são elas: ciência, com 112 ocorrências, e campo, com 14 ocorrências. Esses indicadores temáticos aproximam-se graças ao grau de conexidade entre eles, e também, ao termo informação presente nos documentos. Como o principal termo composto adotado na busca desses documentos foi a "ciência da informação", justifica-se observar o maior grau de conexidade entre as palavras "informação" e "ciência", nesta análise, de 67. 4- O quarto agrupamento observado, em azul escuro, apresenta 2 termos que se relacionam de forma mais intensa, tanto no aspecto de frequência quanto no de coocorrência. (inteligência com 22 ocorrências e competitivo com 18 ocorrências) se apresentam como palavras que ocorrem de maneira muito frequente na descrição dos objetos utilizados nas análises. Percebe-se que esses termos ocorrem frequentemente pois as pesquisas em Ciência da Informação, envolvem, em muitos momentos, os "inteligência competitiva" como objeto principal das análises. 5- Posteriormente, identifica-se o agrupamento mais relevante, que se conecta com "Informação".

Em verde claro, observa-se (Ciência, com 95 ocorrências e campo com 14 ocorrências). Esses termos são os mais relevantes, dentro da análise de similitude, pois a palavra "Ciência" apresenta um índice de conexidade, no valor de 56 em relação à palavra "Informação". Sendo assim, muitos artigos tratam do Campo da CI como área estudo e análise, essas palavras são frequentemente citadas e justificam os valores identificados neste agrupamento. 6- A seguir, observa-se o agrupamento representado pelas palavras (científico com 64 ocorrências, publicação com 12 ocorrências e produção com 21 ocorrências), na cor lilás, essas palavras conectam-se fortemente pois são comumente utilizadas como locução nominal "produção científica" e "publicação científica", nos artigos observados. Esses termos compostos são áreas de estudo da CI e contemplam a área o campo de conhecimento dos EMI. 7-

Posteriormente, identificam-se 2 termos que interagem com a palavra "Informação", (área com 40 ocorrências e leitura com 11 ocorrências). Esses termos possuem relação temática com Informação, pois a área temática em que atuam, principalmente a "área da ciência da informação" e a "leitura e informação" são áreas importantes de estudo da CI. 8- E por fim, ressalta-se o último agrupamento de palavras, com os termos (pós, com 14 ocorrências e graduação, com 12 ocorrências). Esses termos, seguindo a mesma lógica, conectam-se de forma intensa, entre si, ocorrendo com certa frequência em orações devido à sua forte conexão e relação com relação à área da pós-graduação e suas relações com a CI.

Em seguida, verifica-se o gráfico 14 de frequência dos descritores adotados pelos indexados, no *corpus A*, esse gráfico reflete a frequência absoluta dos termos indexados pelos indexadores. Percebe-se que "ciência da informação", "tecnologias da informação e comunicação", "sociedade da informação", são os descritores mais frequentes dentro do conjunto de palavras observado. Isso indica que os indexadores seguiram o alinhamento temático na atribuição de palavras do conjunto observado. Nesse contexto, esse gráfico comprova e justifica as palavras observadas na análise de similitude e no gráfico de frequência das palavras isoladas. No mais, ao observar os termos em sua configuração composta, nota-se que o destaque para "Informação" vem de sua transversalidade, estando presente em várias palavras-chave do *corpus* analisado. Do mesmo modo, Ciência, tem seu significado atrelado à Ciência da Informação, sendo assim, admite-se que estes termos caracterizam o núcleo base da área, servindo ao propósito de caracterizá-la. Assim, importa observar com mais atenção os termos que as sucedem, como por exemplo, "Tecnologias da Informação e Comunicação", "Sociedade da Informação" e "Gestão da Informação", enquanto assuntos mais expressivos.

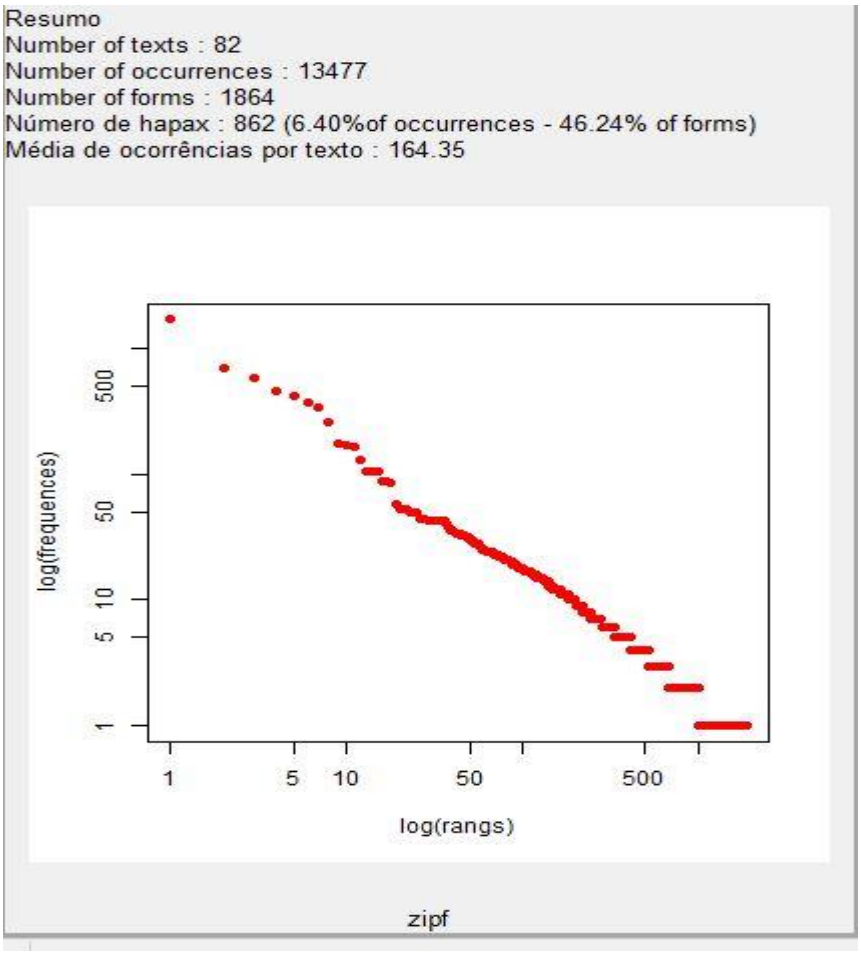
Gráfico 14 - 25 descritores mais frequentes (palavras isoladas/compostas) - Indexadores Manuais – *corpus* A



Fonte: dados da pesquisa (2021).

Adiante, analisaram-se as estatísticas dos indexadores manuais, estruturados na configuração de dispersão presente no gráfico 15.

Gráfico 15 - Estatística dos Indexadores Manuais – *corpus* B

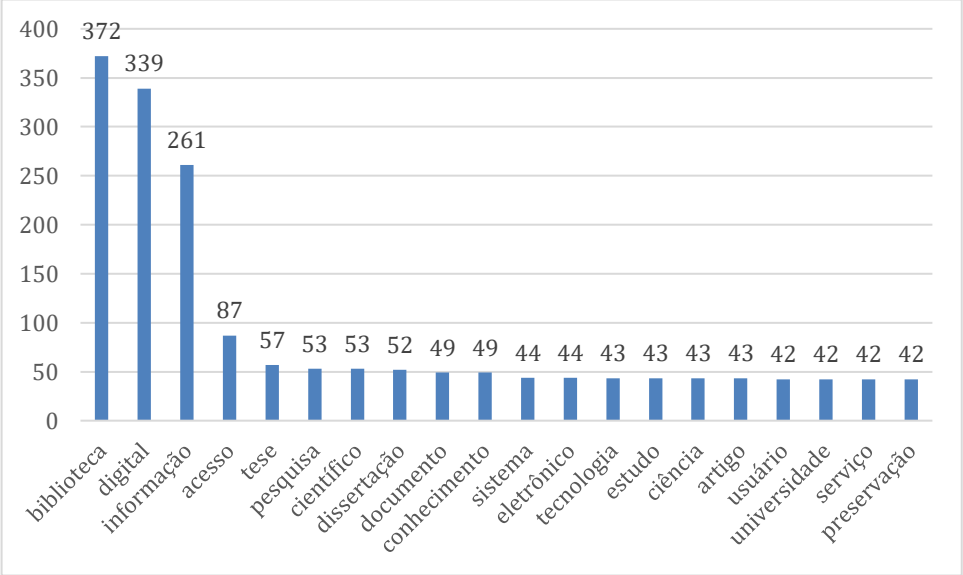


Fonte: dados da pesquisa (2021).

No Gráfico 15, percebe-se a existência de 82 documentos analisados, contendo 13477 ocorrências de termos para 1864 tipos de palavras. Entretanto, cerca de 6,4% das ocorrências, ou 862 palavras, aparecem uma única vez nos documentos, para 46,24% das formas, ou seja, quase metade das palavras utilizadas são citadas uma única vez, enquanto a outra metade é citada 93,6% das vezes. Essa relação corrobora com a aplicação da lei de Zipf e justifica o porquê de poucas palavras representarem o conjunto total analisado.

Em seguida, é observado o gráfico 16 de frequência das palavras isoladas do *corpus* B, com o conjunto de termos indexados pelos indexadores manuais. Foram selecionados os 20 termos mais frequentes do *corpus* com base na sua ocorrência dentro do conjunto textual. Percebe-se as palavras “biblioteca” e “digital” como as mais frequentes dentro do conjunto observado.

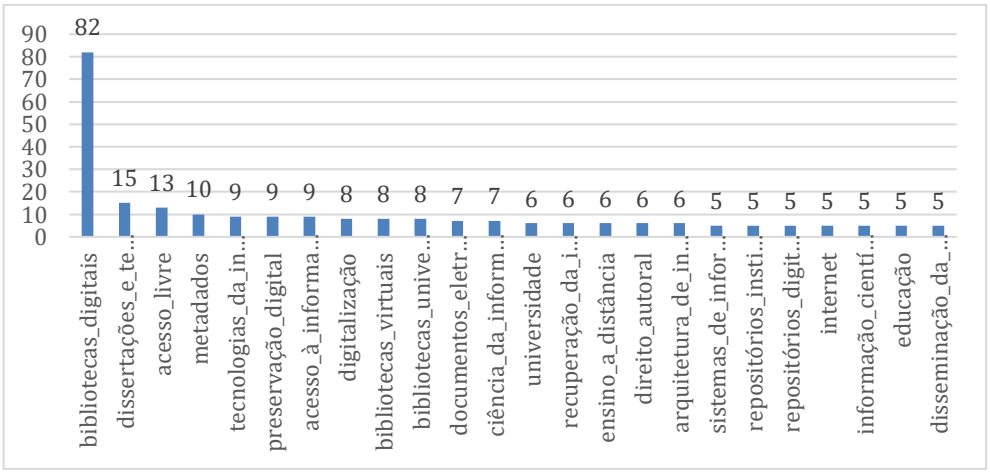
Gráfico 16 - 20 palavras mais frequentes (palavras isoladas) - Indexadores Manuais – *corpus* B



Fonte: dados da pesquisa (2021).

A análise de palavras unitárias pode ser complementada pelos termos compostos do próximo gráfico. Nestas ilustrações fica evidente a representatividade das bibliotecas digitais, que podem ser entendidas como conjuntos de fontes eletrônicas e serviços técnicos associados para a criação, pesquisa e uso da informação, que possibilitam uma extensão e um aumento do armazenamento da informação e dos sistemas de recuperação de informação, manipulando dados digitais em qualquer meio (texto, imagens, sons, imagens dinâmicas e estáticas) em redes distribuídas de trabalho (BORGMAN, 1996). Desta feita, assume-se que o *corpus* estudado foi produzido por um grupo especializado neste assunto, dado seu amplo destaque nas estatísticas.

Gráfico 17 - 24 descritores mais frequentes (palavras isoladas/compostas) - Indexadores Manuais – *corpus* B

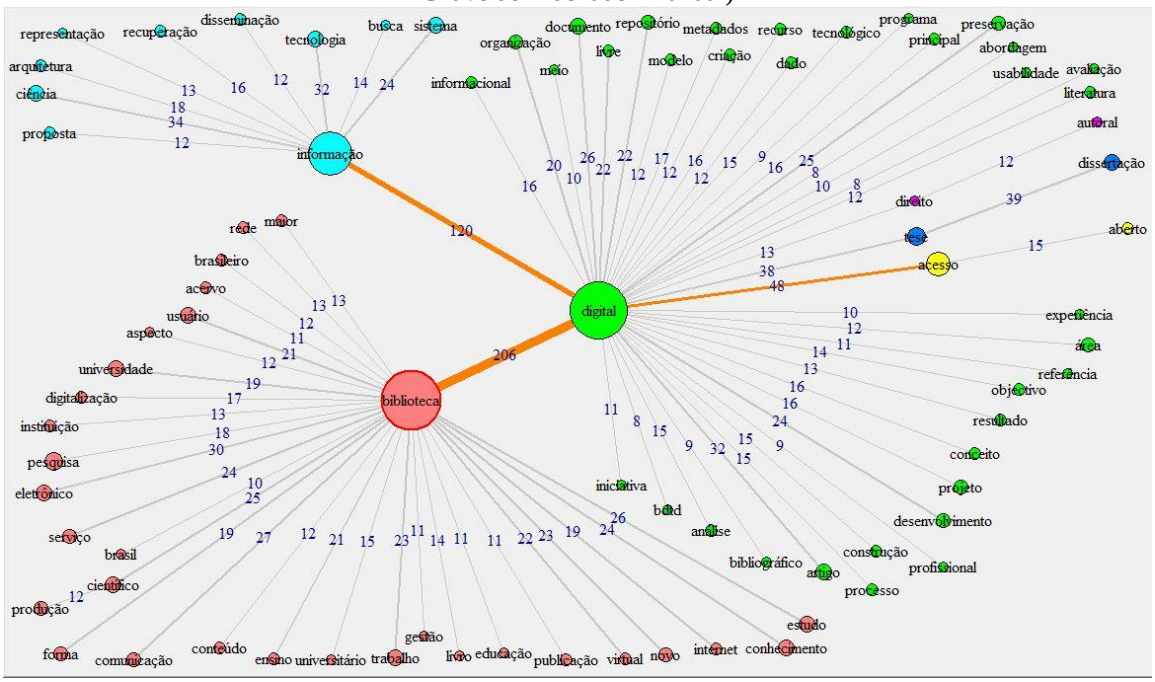


Fonte: dados da pesquisa (2021).

O gráfico 17 confirma os achados do gráfico 16, e de modo complementar, reforça uma interessante descoberta sobre o *corpus*. Ao analisar os termos principais, vê-se um grande enfoque na área de tecnologia, presente nas bibliotecas digitais, nos metadados, nas TICs, na preservação digital, na digitalização e nas bibliotecas virtuais. Tais indicadores corroboram com a visão de Roza (2018), quando afirma que as TICs são responsáveis, em parte, pelas diversas transformações na sociedade da informação, levando alguns autores a adotarem abordagens tecnológicas no campo informacional.

Em seguida, na figura 15, é possível analisar o gráfico de análise de similitude com as 83 palavras mais frequentes, com ocorrência mínima de 16 unidades dentro do *corpus* B. Neste sentido, foram identificados os principais eixos temáticos que se relacionam com a "Biblioteca Digital". Na investigação quantitativa foi identificado que os termos "Biblioteca" e "Digital" continuam sendo as palavras com maior índice de frequência, dentro dos documentos, apresentando 372 e 339 ocorrências. Assim, com base na teoria dos grafos, são identificadas as ligações que conectam os termos de acordo com o seu grau de proximidade e conexidade, principalmente, com os termos: "Informação" e "Acesso" (vide Figura 15).

Figura 15- Análise de Agrupamentos (Clusters) do *corpus* B com os Metadados (Título, Resumo e Palavras-Chave do Indexador Manual)



Fonte: dados da pesquisa (2021).

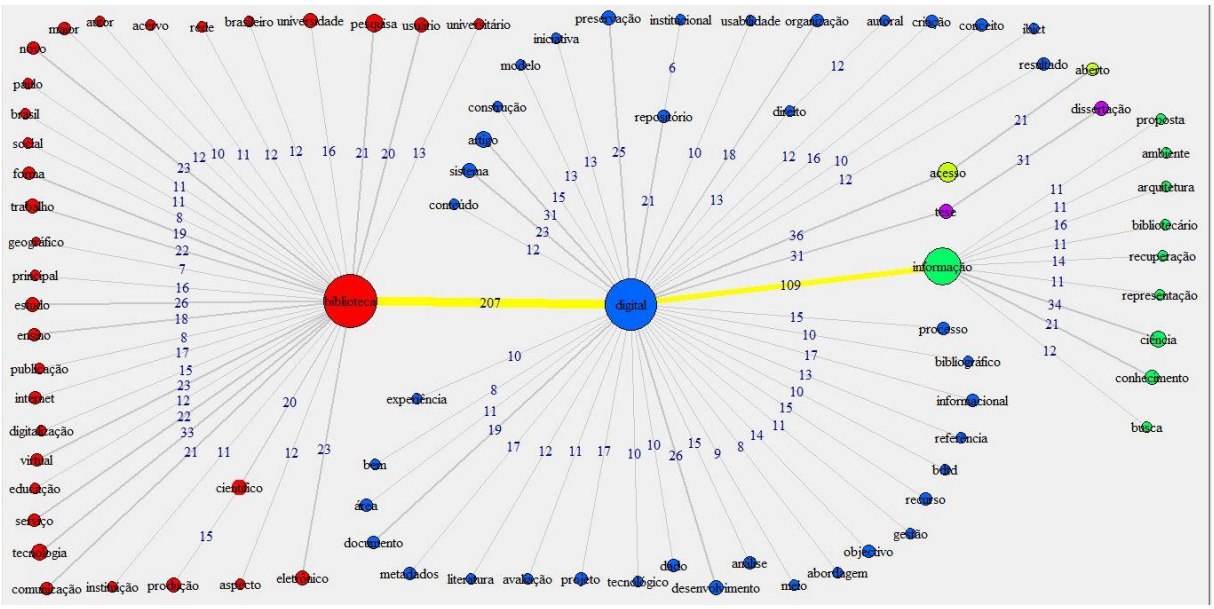
Verifica-se a existência de 3 grandes agrupamentos de palavras e 3 menores, mais emergentes, tal característica ratifica a relação da conjuntura temática do *corpus* estudado,

pois os 82 trabalhos analisados compõem o campo da "Biblioteca Digital". Portanto, após a observação desses grandes grupos foi possível perceber a clusterização das palavras, objetivando apontar os termos que se agrupam por critérios relacionais e de similitude no grafo. Dito isto, perceberam-se agrupamentos compostos por palavras distintas, a saber: 1- O primeiro agrupamento, na cor azul piscina, possui 10 palavras identificadas nos círculos que se conectam entre si com a palavra "Informação" com 261 ocorrências. 2- O segundo agrupamento, está representado na cor vermelho claro com cerca de 30 palavras conectadas com a palavra "Biblioteca" com 372 ocorrências. 3- O terceiro agrupamento observado, em verde, apresenta 35 palavras que se relacionam de forma mais intensa, tanto no aspecto de frequência quanto no de coocorrência com a palavra digital. 4- Posteriormente, o agrupamento em roxo, apresenta 2 termos ("Direito" com 21 ocorrências e "Autoral" com 17 ocorrências). 5- Em seguida, observa-se o agrupamento representado pelas palavras (tese com 57 ocorrências e dissertação com 52 ocorrências), na cor azul escura, essas palavras conectam-se fortemente pois são comumente utilizadas como locução nominal "teses e dissertações", identificadas no campo científico principalmente como objetos de análise em pesquisas do campo da CI. 6- E, finalmente, no sexto agrupamento, observam-se as palavras ("Acesso", com 87 ocorrências e "Aberto", com 24 ocorrências), esses termos são representativos e ocorrem de maneira frequente, tanto em grau de proximidade e de coocorrência pelo campo de atuação do "Acesso aberto", "Acesso a dados abertos", "Acesso aberto à informação", etc. Além disso, são áreas relevantes e com diversos estudos[40] dentro da CI. Diante disso, os principais indicadores temáticos identificados são devidamente representados pelo campo de conhecimento que atuam. Os termos possuem relação com a área das "bibliotecas digitais" e são reiteradamente utilizados na área.

A Posteriori, elaborou-se uma análise similar no *corpus* B, observando o grau de conexidade entre as palavras dos metadados observados com as palavras-chave do autor, que representam os 82 artigos investigados. Ressalta-se que para a obtenção do resultado alcançado na figura 16, foram utilizados os 87 termos mais frequentes e relevantes do *corpus*, de uma ocorrência que vai de 15 a 387 de cada palavra, da menos frequente para a mais frequente.

[40] SILVA, Terezinha Elisabeth da; ALCARÁ, Adriana Rosecler. Acesso aberto à informação científica: políticas e iniciativas governamentais. **Informação & Informação**, v. 14, n. 2, p. 100-116, 2009.

Figura 16 - Análise de Agrupamentos (Clusters) do *corpus* B com os Metadados (Título, Resumo e Palavras-Chave do Autor)



Fonte: dados da pesquisa (2021).

Nesse tipo de visualização é possível identificar 2 polos centrais (“Biblioteca”, em vermelho e “Digital”, na cor azul). Ambas as palavras, apresentam-se com um certo grau de centralidade e representatividade temática. Na análise de frequência foi identificado que o termo “Biblioteca” é a palavra mais frequente, apresentando 381 ocorrências, e a palavra “Digital” é a segunda mais frequente, com 357 ocorrências. Portanto, percebe-se uma grande espessura das arestas que conectam as palavras observadas, com um índice de 207 no grau de coocorrência, indicando uma grande relação entre esses dois termos mais frequentes. (vide Figura 16).

Verifica-se a existência de 32 palavras, na cor vermelha, mais à esquerda, que aparecem ligadas ao termo “Biblioteca”, identificando assim uma relação mais forte com essas palavras. Em relação à palavra “Digital”, percebe-se a existência de 39 palavras, na cor azul, que estão mais intensamente conectadas a este termo. Outro fenômeno interessante, é a existência de 3 grupos de palavras mais à esquerda, sendo um deles com a palavra “Informação”, com 229 ocorrências. Apesar do alto índice de ocorrência da palavra informação, ela não é a palavra mais importante desse *corpus*, diferentemente do que ocorreu no *corpus* A, quando “Informação” sempre surgia como o termo principal mais relevante. Isso ocorre graças à conjuntura temática do *corpus* estudado, os 82 artigos analisados neste *corpus* pertencem ao campo da “Biblioteca Digital”, pois esta palavra-chave foi utilizada no campo de busca para encontrar os 82 artigos presentes no *corpus* B. Nesse âmbito, apesar da grande

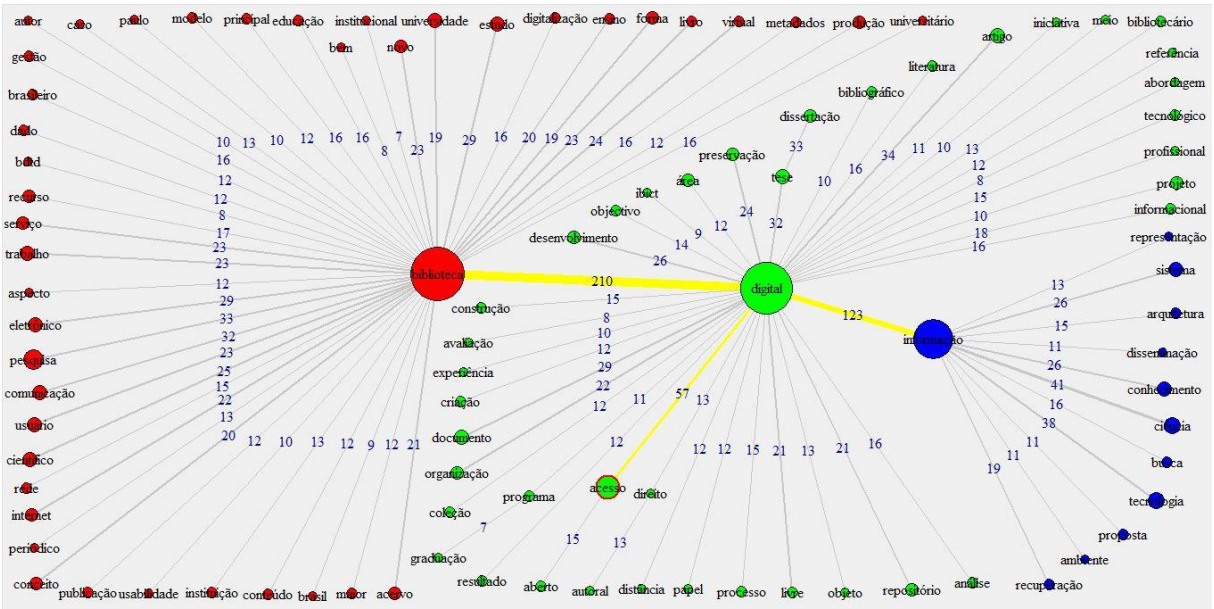
importância e relação existente entre a palavra “Informação” e “Digital”, essa locução nominal não é a mais frequente no *corpus*.

A partir desse ponto de vista, é importante analisar o grau de proximidade das palavras agrupadas nos 3 agrupamentos mais emergentes. Identificam-se os seguintes agrupamentos, de acordo com a ordem de proximidade ao termo “Digital”: 1- O primeiro agrupamento, na cor amarelo escura, possui 2 palavras identificadas (“Acesso” com 75 ocorrências e “Aberto” com 36 ocorrências). Essas palavras ocorrem, com maior frequência, próximas à palavra “Digital”. Ambas são representativas, em maior ou menor grau, porém, a palavra “Acesso” possui maior relevância e representatividade, no contexto observado nesse *corpus*. Essa visibilidade deve ocorrer, pois muitos artigos são voltados para área de “Acesso aberto”, “Acesso à Informação”, “Acesso Digital” ou “Acesso à biblioteca digital”. Portanto, a palavra “Acesso” surge como um termo relevante dentro do contexto da CI. 2- O segundo agrupamento está representado na cor lilás com 2 palavras conectadas (“Tese”, com 53 ocorrências, e “Dissertação”, com 46 ocorrências), essas palavras se relacionam de forma mais próxima entre elas, possivelmente, devido aos estudos sobre os repositórios de dados e institucionais que são, na verdade, bibliotecas digitais.

Nesses repositórios estão contidos dos documentos digitais de Teses e Dissertações de pesquisadores, formados nessas instituições. Nesse contexto, essa correlação pode ser justificada, pois as locuções “Teses e Dissertações” e “Bibliotecas digitais de Teses e Dissertações” são habitualmente utilizadas para descrever essas áreas de interesse. 3- O terceiro agrupamento observado, em verde, apresenta 9 termos que se relacionam de forma mais intensa, com a palavra Informação. Neste caso, as palavras (“Proposta”, com 18 ocorrências, “Ambiente”, com 15 ocorrências, “Arquitetura”, com 23 ocorrências, “Bibliotecário”, com 16 ocorrências, “Recuperação”, com 17 ocorrências, “Representação”, com 16 ocorrências, “Ciência”, com 43 ocorrências, “Conhecimento”, com 47 ocorrências e “busca”, com 22 ocorrências) se apresentam como palavras que se conectam, de maneira muito frequente, com a palavra “Informação”, que por sua vez, conecta-se à palavra digital. Diante disso, percebe-se a existência profunda dessas palavras com diversas áreas temáticas que envolvem: proposta de informação digital, ambiente de informação digital, arquitetura de informação digital, bibliotecário de informação digital etc. Justificam-se essas relações pois as áreas temáticas convergem para a “biblioteca digital”, percebendo-se como um campo de pesquisa forte, dentro da CI.

Seguidamente, a figura 17, a seguir, foi feita com 97 termos mais frequentes e relevantes do *corpus* B (substituindo as palavras-chave do autor pelos descritores atribuídos pelo MAUI), de uma ocorrência que vai de 15 a 411 de cada palavra, da menos frequente para a mais frequente. Essa figura representa a visualização no formato de gráfico de *co-words*, das palavras utilizadas nos metadados identificados a seguir. *A priori*, observa-se uma sútil modificação nos agrupamentos de palavras, percebendo-se apenas 3 grupos de palavras identificáveis, perante os 5 grupos anteriormente identificados. Nota-se, principalmente, o aumento no grau de conexão entre as palavras "Acesso" e "Digital", ampliando de índice de 36 para 57. Isso demonstra que o *software* Maui indexou muito mais o termo "Acesso" próximo à palavra "Digital". O mesmo cenário se apresenta para a relação entre os termos "Tese", "Dissertação" e Digital". "Para essas palavras, percebe-se possíveis ligações entre os termos e assuntos que coocorrem, proporcionando, desta forma, a criação de um mapa do conhecimento" (FRANCO; FARIA, 2019, p. 90).

Figura 17 - Análise de Agrupamentos (Clusters) do *corpus* B com os Metadados (Título, Resumo e Palavras-Chave do Maui)



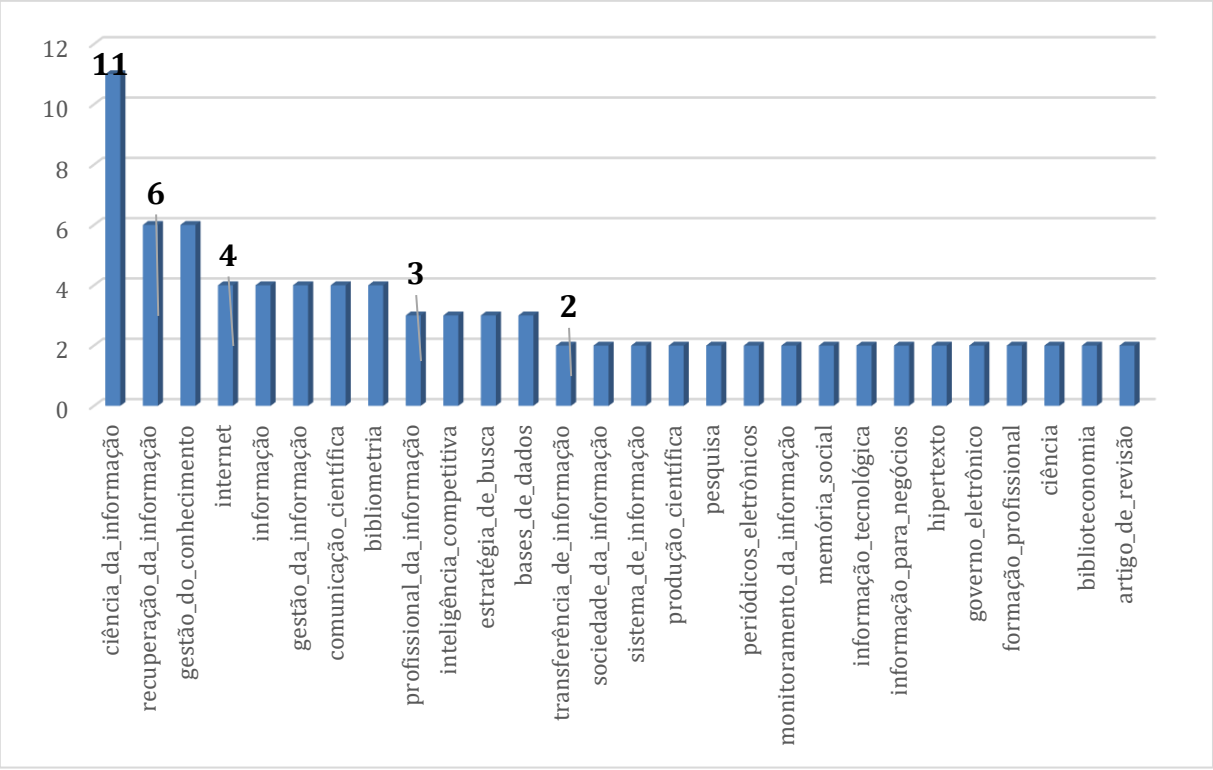
Fonte: dados da pesquisa (2021).

Neste sentido, constata-se a existência de 43 palavras, na cor vermelha, mais à esquerda, que aparecem ligadas ao termo "Biblioteca", identificando assim uma relação mais forte com essas palavras, e cerca de 40 palavras conectadas ao vocábulo "Digital", na cor verde, que estão mais intensamente conectadas a este termo. Outro fenômeno interessante, é a existência de apenas 1 grupo de palavras mais à esquerda, em azul, possuindo o termo central "Informação", com 229 ocorrências, termos, estes, mais relevantes do grupo. Observa-se com isso, que diferentemente da análise de similitude da figura 16, a figura 15 apresenta apenas 3

grandes grupos de palavras, pois 2 deles foram absorvidos pelo agrupamento em verde. Esse fenômeno indica que o Maui indexou muito mais as palavras-chave (Tese, dissertação, Acesso e Aberto) de forma coocorrente com a palavra "digital", tornando-as assim, pertencentes ao agrupamento em torno do vocábulo "digital". E, concomitantemente, como foi afirmado, não obstante o alto índice de ocorrência da palavra "informação", ela não é a palavra mais importante desse *corpus*, diferentemente do que ocorreu no *corpus* A, quando "Informação" sempre surgia como o termo principal mais relevante.

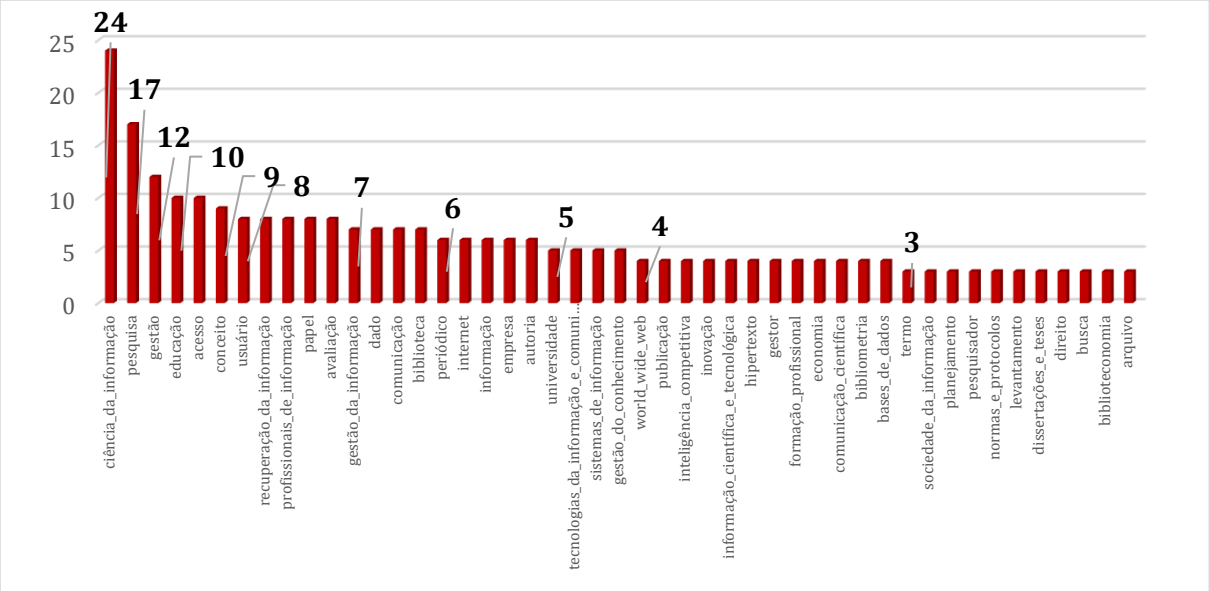
Os resultados obtidos nas etapas 4 e 5 do fluxograma são muito relevantes pois demonstram a importância das palavras-chave e sua comparação terminológica. A opção por elaborar gráficos de frequência de palavras-chave teve como objetivo principal apresentar a relevância dos descritores, pois como afirmaram Franco e Faria (2019, p. 92), as palavras-chave são termos indexadores que indicam um breve resumo do conteúdo documental, além disso, elas sintetizam o texto permitindo uma visão simplificada do documento auxiliando na descrição dos assuntos retratados. Para identificar os termos compostos, elaborou-se o Gráfico 18 e o Gráfico 19.

Gráfico 18 - Frequência de Palavras-chave atribuídas manualmente do *corpus* A dos Autores



Fonte: dados da pesquisa (2021).

Gráfico 19 - Frequência de descritores do Maui para o *corpus* A



Fonte: dados da pesquisa (2021).

Após a visualização dos resultados obtidos nos gráficos 18 e 19, é possível identificar uma variação na quantidade de palavras atribuídas em cada método. É perceptível uma mudança tanto quantitativa, quanto na escolha de determinados termos que não foram previamente atribuídos, neste sentido, o Maui, por meio do método de aprendizagem de máquina por vocabulário controlado, conseguiu atribuir "Ciência da Informação" como um descritor, 13 unidades a mais, "Recuperação da Informação", 3 unidades a mais, "Internet", 2 unidades a mais, "Informação", 1 unidades a mais, "Gestão da Informação" 3 unidades a mais. Dois descritores se comportaram de forma atípica e não tiveram aumento de atribuição pelo Maui, "Comunicação Científica", que manteve o valor, e, de maneira oposta, "Gestão do Conhecimento" que foi atribuída menos 1 vez, pelo sistema. Nesse cenário, podem ser percebidos e analisados os diversos termos propostos na indexação. Para facilitar essa comparação de resultados, elaborou-se a tabela 06, a seguir.

Tabela 6 - Análise Comparativa entre os 7 termos manuais mais frequentes do *corpus* A

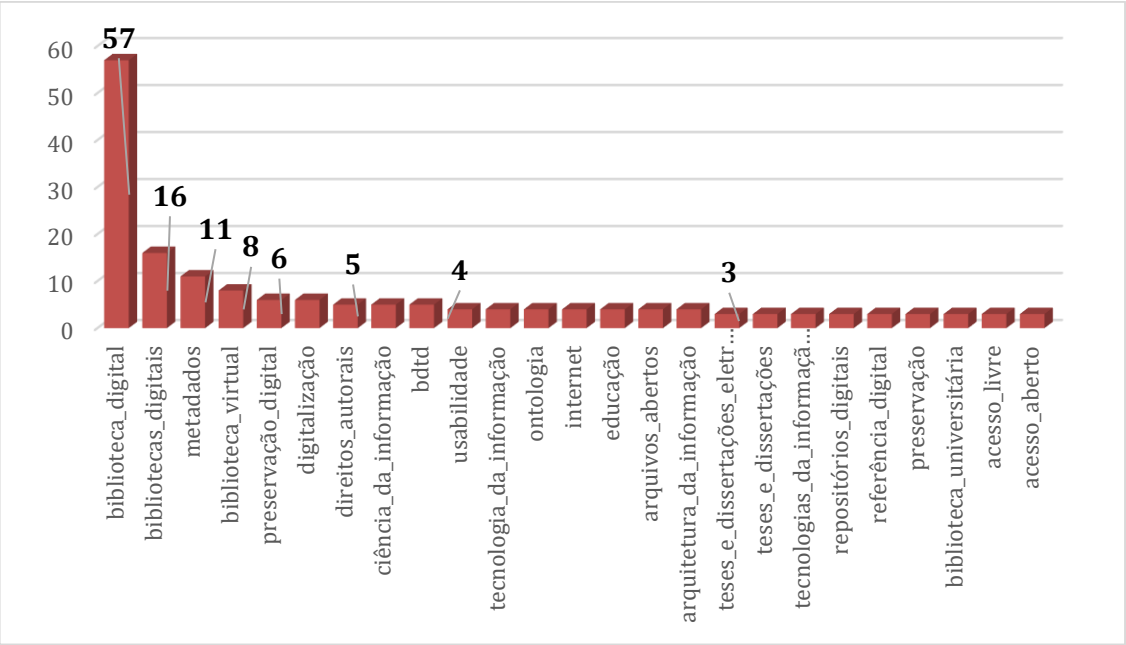
FREQUÊNCIA DE PALAVRAS	CIÊNCIA DA INFORMAÇÃO	RECUPERAÇÃO DA INFORMAÇÃO	GESTÃO DO CONHECIMENTO	INTERNET	INFORMAÇÃO	GESTÃO DA INFORMAÇÃO	COMUNICAÇÃO CIENTÍFICA
MANUAL	11	6	6	4	4	4	4
MAUI	24	9	6	6	6	7	4
VARIAÇÃO	13	3	-1	2	2	3	0

Fonte: dados da pesquisa (2021).

A partir da Tabela 7 é possível observar a diferença existente entre cada tipo de atribuição. Optou-se por priorizar as 7 palavras-chave mais frequentes, atribuídas pelos autores dos artigos. Nesse quadro, foi possível identificar que o Maui priorizou a indexação de outros termos mais frequentes, além de outros vocábulos que não necessariamente foram utilizados pelos indexadores, como: educação, conceito, usuário, dado, empresa, dentre outros.

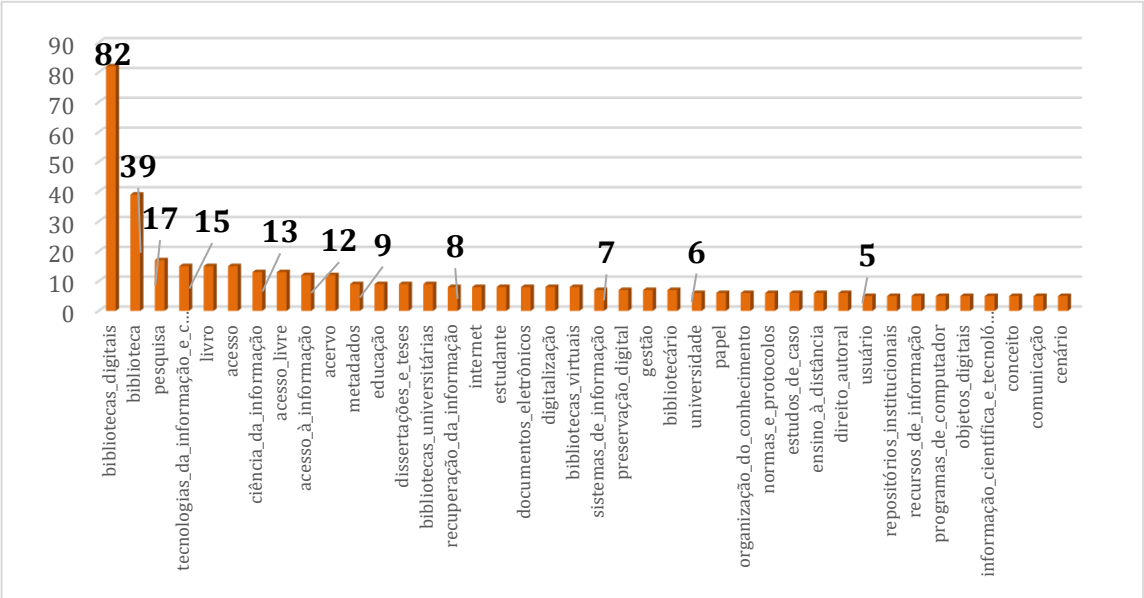
De maneira semelhante, após a visualização dos resultados obtidos no *corpus* A, foi realizada a análise de palavras-chave no *corpus* B, como pode ser visualizado nos 2 gráficos, em seguida (Gráfico 20 e 21).

Gráfico 20 - Frequência de Palavras-chave manuais do *corpus* B



Fonte: dados da pesquisa (2021).

Gráfico 21- Frequência de Palavras-chave do Maui do *corpus* B



Fonte: dados da pesquisa (2021).

Após a visualização de ambos os gráficos, comparando as tipologias de termos atribuídos manualmente e automaticamente, é possível identificar uma variação na quantidade de palavras atribuídas em cada método. É perceptível uma mudança tanto quantitativa, quanto na escolha de determinados termos que não foram previamente atribuídos, neste sentido, o Maui, por meio do método de aprendizagem de máquina por vocabulário controlado, conseguiu atribuir "Bibliotecas Digitais" como um descritor, 9 unidades a mais, "Metadados", 2 unidades a mais, "Preservação digital", 1 unidade a mais, "Digitalização", 2 unidades a mais, "Direitos autorais", 1 unidade a mais e "Ciência da Informação", 8 unidades a mais. Apenas 1 descritor, dos 7 descritores mais frequentes atribuídos manualmente, manteve-se na quantidade e não teve um aumento de atribuição pelo Maui, que foi, "Biblioteca Virtual", que manteve a quantidade de ocorrências no *corpus*. Uma observação importante com relação à comparação entre a frequência de atribuição dos termos, sugere que o Maui identificou que os documentos indexados possuíam maior relação com à temática das "Bibliotecas Digitais", assim sendo, o sistema foi capaz de relacionar a temática principal dos documentos a uma quantidade maior de artigos. Dessa forma, podem ser avaliados e estudados os diversos termos propostos na indexação. Para facilitar essa comparação de resultados, elaborou-se a tabela 07, a seguir.

Tabela 7 - Análise Comparativa entre os 7 termos manuais mais frequentes do *corpus* B

FREQUÊNCIA DE PALAVRAS	BIBLIOTECAS DIGITAIS	METADADOS	BIBLIOTECA VIRTUAL	PRESERVAÇÃO DIGITAL	DIGITALIZAÇÃO	DIREITOS AUTORAIS	CIÊNCIA DA INFORMAÇÃO
AUTOR	73	11	8	6	6	5	5
MAUI	82	9	8	7	8	6	13
VARIAÇÃO	9	2	0	1	2	1	8

Fonte: dados da pesquisa (2021).

A partir da Tabela 08 é possível observar a diferença existente entre cada tipo de atribuição. Optou-se por priorizar as 7 palavras-chave mais frequentes, atribuídas manualmente. Nessa circunstância, foi possível identificar que o Maui priorizou a indexação de outros termos mais frequentes, além de outros vocábulos que não necessariamente foram utilizados pelos autores, como: biblioteca, pesquisa, livro, acervo, dentre outros.

Com base nos resultados obtidos, fundamentados nos estudos que envolvem, tanto a aplicação das metodologias da IA, quanto dos EMI, elaboraram-se quatro inferências principais sobre os resultados alcançados nesta análise:

1- Inferência 1 -> *Corpus* A: a produção científica apresentou enfoque em assuntos ligados à gestão e atuação profissional, sendo isto visível na disposição das palavras-chave.
2- Inferência 2 -> *Corpus* B: a produção científica demonstrou viés associado a tecnologias digitais, comportamento explícito na constelação de termos.
3- Inferência 3 -> os índices de indexação e revocação dos *corpora* encontram-se adequados às expectativas encontradas nas bibliografias científicas especializadas.
4- Inferência 4 -> ficou constatada a eficácia do percurso divido em etapas, primeiro com a adoção dos aportes da IA, e, depois, dos aportes dos EMI. Sendo assim, analisaram-se os termos indexados e suas relações de coocorrência.

Portanto, os resultados apresentaram gráficos estatísticos, gráficos de análise de similitude, gráficos de frequências de palavras e tabelas de métricas da IA. Além disso, foram inseridos apêndices adicionais (vide Apêndice C, D e E) com o propósito de facilitar o entendimento e a compreensão do que foi elaborado.

Por fim, na seção seguinte, foram descritas as principais considerações sobre a eficiência do percurso metodológico proposto. Dada a natureza do trabalho científico, que possui um espaço próprio para esse conteúdo, as informações acerca disto, encontram-se na seção de Conclusão.

5 CONCLUSÃO

Esta pesquisa de doutorado objetivou propor um percurso metodológico para a formulação de IT de informação científica, a partir da relação teórica e metodológica entre a IA e os EMI. Para isso, realizaram-se: a identificação dos pontos de integração teórica e metodológica entre a IA e os EMI, constatando relação de complementariedade, especialmente no que tange aos aportes da teoria dos grafos, materializada na análise de similitude e coocorrência; o delineamento da trajetória metodológica de formulação de IT de informação científica, valendo-se de modelagem de processos, por meio da adoção do fluxograma, com o intuito de prover subsídios para replicação em estudos futuros; e a implementação das análises e processamentos dos corpora definidos na pesquisa, legitimando os resultados obtidos no percurso metodológico.

Ao final, percebeu-se a busca por uma integração teórico-metodológica nos processos de elaboração dos Indicadores Temáticos, com os aportes da Indexação Automática (IA) e dos Estudos Métricos da Informação (EMI); verificou-se a utilidade desse processo ao entender que as métricas da IA, associadas ao software adotado, correspondem aos valores compatíveis a bons índices de Indexação, demonstrando a importância da utilização dessa metodologia, que pode acelerar o processo da descoberta de conhecimento.

Ao serem extraídos os conceitos principais de um determinado conjunto documental, de forma automatizada, o processo se tornou mais eficiente, possibilitando um acréscimo da agilidade na identificação de palavras relevantes, bem como acelerando a representação temática com a construção dos IT propostos.

A metodologia adotada pode impulsionar avanços na diminuição do trabalho manual a ser realizado, reduzindo custos e promovendo uma evolução na implementação de políticas de indexação mais eficientes. Neste sentido, as vantagens proporcionadas no percurso adotado na IA causam uma elevação no nível de representatividade visual dos Indicadores temáticos obtidos.

Após a finalização do processo, com tal nível de representação, foi possível visualizar o alto grau de correlação terminológica entre os principais conceitos mais influentes, num determinado campo científico ou domínio, e as relações conceituais primordiais, identificadas pela ferramenta.

Dessa forma, os resultados obtidos pelo percurso adotado podem servir de subsídio tanto na definição de metas para produção de conhecimento quanto na elaboração de diretrizes que possam fomentar a produção acadêmica em áreas prioritárias. Portanto, o

desenvolvimento de pesquisas que incentivem a descoberta de conhecimento, em grandes volumes documentais, pode ser financiado por agências de fomento nacional, como a CAPES e o CNPq, ou estadual, a exemplo da FACEPE, com o intuito de subsidiar e estimular o crescimento deste campo.

Destarte, a fim de apresentar as principais considerações sobre a tese, aqui desenvolvida, revelam-se as principais considerações sobre o primeiro objetivo específico, cujo propósito foi: identificar os elementos de integração teórico e metodológica, propostos neste documento.

Tal objetivo foi alcançado, ao se discutirem os fundamentos conceituais e empíricos da cartografia temática, bibliometria temática e IA, aplicada ao mapeamento bibliométrico. Os aportes teóricos definidos, nas pesquisas apresentadas nestas áreas, foram basilares para corroborar com os conceitos basilares trazidos pelos EMI e pela IA. Assim, o campo científico foi devidamente contemplado, obtendo todos os elementos de discussão pertinentes ao estudo proposto.

No que se refere ao segundo objetivo específico, que almejou delinear a trajetória metodológica para formulação de indicadores de temáticas, pode-se apontar que ele foi devidamente atendido, ao apresentar todas as etapas envolvidas no capítulo 3 dos procedimentos metodológicos. Dentre as etapas, destacam-se os fluxogramas da IA e do Estudo Métrico, que tinham como finalidade explicitar os processos envolvidos.

Em relação ao terceiro objetivo, houve a implementação das análises e processamentos dos corpora, definidos na pesquisa, percebendo-se que os principais índices da indexação se comportaram dentro dos valores esperados.

Quanto ao problema de pesquisa, questionou-se sobre a configuração das relações entre a IA e os EMI para a formulação metodológica de IT de informação científica, obtendo-se as seguintes respostas: perceberam-se as vantagens e potencialidades da aplicação conjunta dos EMI e da IA, sobretudo no que concerne à produção de métodos de visualização de IT; identificaram-se os índices e medidas associadas ao processo de IA por atribuição; analisaram-se os principais termos relevantes nas análises de similitude e de coocorrência das palavras-chave; compararam-se os principais termos atribuídos pelo autor e as palavras indexadas pelo software Maui; e, por fim, fez-se a avaliação para verificar se os resultados obtidos pelo Maui, após a IA, oferecem segurança na geração de bons resultados nas visualizações com o Iramuteq.

Com relação aos softwares adotados, notaram-se as vantagens e desvantagens cardinais, que puderam ser observadas, ao longo da estruturação do percurso metodológico:

a) Com relação ao software Maui, percebem-se as seguintes vantagens: (i) realiza a indexação automática por atribuição, através de vocabulário controlado; (ii) utiliza a aprendizagem de máquina para aprender o vocabulário da área; (iii) é um software gratuito com código aberto. E tais desvantagens: (i) existe uma dificuldade no processo de execução do software, com relação à experiência do usuário, pois é necessário acessar o prompt do comando no Windows, para executar os comandos necessários; (ii) é requisitado determinado conhecimento básico prévio na área técnica de informática, para facilitar o processo de execução da ferramenta.

b) No que diz respeito ao software Iramuteq, notam-se as seguintes vantagens: (i) implementa uma análise estatística textual, com a adoção de leis bibliométricas já consagradas; (ii) possui biblioteca disponível para língua portuguesa, possibilitando o processo de execução em documentos neste idioma; (iii) é um software gratuito com código aberto, tal como o Maui. Em relação às desvantagens do software Iramuteq, enumeram-se: (i) existe uma dificuldade no processo de execução do software, com relação à experiência do usuário, pois é necessário aprender as etapas previamente, para executar os comandos necessários; (ii) também é preciso ter conhecimento básico prévio na área técnica de informática, para facilitar o processo de execução da ferramenta.

Uma consideração que pode ser feita, a partir dos pontos destacados acima, é que, apesar das desvantagens identificadas em ambos os sistemas, percebe-se uma eficiência no processo de execução, com os resultados bem definidos.

Contudo, o desenvolvimento da pesquisa apresentou algumas limitações, dentre elas, ressalta-se que a parte final da pesquisa foi realizada num contexto pandêmico, ocasionado pelo novo coronavírus (Sars-CoV-2), responsável por modificar as relações e os modos de produção em todo o mundo, implicando em danos emocionais e físicos, e, infelizmente, no adoecimento e perda de pessoas importantes.

Sobre a pesquisa, especificamente, admite-se que o algoritmo de processamento e visualização dos termos poderia ter realizado contextualizações mais densas, que, por exemplo, permitissem, a detecção de termos compostos, ampliando as análises para além das palavras isoladas. A fim de mitigar tal prejuízo, fez-se necessária a geração de análises adicionais com o software Microsoft Excel, apoiando-se na frequência das palavras, com o intuito de manter a estrutura dos termos compostos em alguns gráficos, preservando as estruturas sintática e semântica do léxico presente nos corpora.

Sugere-se, para estudos futuros, a elaboração de uma plataforma online, que possa fazer a indexação automática dos termos autorizados da CI, amparando-se em tesauros

atualizados. Em tom de sugestão, recomenda-se a observação do modelo utilizado pelos Descritores em Ciências da Saúde, da Biblioteca Virtual em Saúde (DeCS/BVS).

Tal iniciativa, baseada no Medical Subject Headings (MeSH), do National Center for Biotechnology Information, U.S., e da National Library of Medicine, U.S., é um padrão consolidado na obtenção de termos autorizados, amparados em vocabulário controlado, sendo aporte para a criação de indicadores temáticos, como, por exemplo, na pesquisa de Sobral (2019).

Além disso, prescreve-se a criação de um padrão tutorial, explicando cada etapa necessária ao desenvolvimento de um modelo de gráficos de análise de similitude, o que pode auxiliar os pesquisadores e publicadores a produzirem seus próprios IT, validando aplicações da IA.

Por fim, pretende-se, ainda, adotar outros softwares para a etapa da IA, como o SISA, por exemplo, em outros corpora e em outras áreas do conhecimento. Intenta-se, também, a utilização do software VOSviewer, para a etapa dos EMI, a fim de comparar a geração de IT e da visualização temática de ambas as etapas adotadas.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. *In*: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 20., 1994, Santiago de Chile. **Proceedings**... Santiago de Chile: Morgan Kaufmann, 1994. p. 487-499.

ALVAREZ, G. R.; CAREGNATO, S. E. A Ciência da Informação e sua contribuição para a avaliação do conhecimento científico. **Biblos**, [*s.l.*], v. 31, n. 1, p. 09-26, 2017. Disponível em: https://periodicos.furg.br/biblos/article/view/5987. Acesso em: 16 out. 2021.

AMIN, A. *et al*. TOP-Rank: A novel unsupervised approach for topic prediction using keyphrase extraction for urdu documents. **IEEE Access**, [*s. l.*], v. 8, p. 212675-212686, 2020. Disponível em: https://ieeexplore.ieee.org/abstract/document/9265205. Acesso em: 16 out. 2021.

ANDERSON, J. D.; PÉREZ-CARBALLO, J. The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. **Information processing & management**, [*s. l.*], v. 37, n. 2, p. 231-254, 2001. Disponível em: https://www.sciencedirect.ez10.periodicos.capes.gov.br/science/article/pii/S0306457300000261?via%3Dihub. Acesso em: 16 out. 2021.

ARAÚJO, C. A. A. Bibliometria: evolução histórica e questões atuais. **Em Questão**, [*s. l.*], v. 12, n. 1, p. 11-32, 2006. Disponível em: http://hdl.handle.net/20.500.11959/brapci/10124. Acesso em: 16 out. 2021.

ARAÚJO, C. A. A. **O que é Ciência da Informação**. Belo Horizonte: KMA, 2018. 132 p.

ARAÚJO, D. A. O. *et al.* Descoberta de Conhecimentos sobre a esquistossomose a partir de documentos científicos utilizando técnicas de mineração de textos. **Pesq. Bras. em Ci. da Inf. e Bi**., João Pessoa, v.11, n.2, p. 173-186. 2016. Disponível em: www.periodicos.ufpb.br/ojs2/index.php/pbcib/article/view/31846. Acesso em: 16 out. 2021.

ARAÚJO, R. F.; FURNIVAL, A. C. M. Comunicação científica e atenção online: em busca de colégios virtuais que sustentam métricas alternativas. **Informação & Informação**, Londrina, v. 21, n. 2, p.68-89, maio/ago., 2016. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27297. Acesso em: 16 out. 2021.

ARONSON, A. R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *In*: **Proc AMIA Symp**, p.17-21, 2001. Disponível em: https://pubmed.ncbi.nlm.nih.gov/11825149/. Acesso em: 16 out. 2021.

ASSEFA, S. G.; RORISSA, A. A bibliometric mapping of the structure of STEM education using co-word analysis. **Journal of the American Society for Information Science and Technology**, [*s. l.*], v. 64, n. 12, p. 2513-2536, 2013. Disponível em: https://onlinelibrary.wiley.com/doi/epdf/10.1002/asi.22917. Acesso em: 16 out. 2021.

BANDIM, M. A. S. **Indexação Automática por Atribuição de Artigos Científicos da Área de Ciência da Informação**. 2017. 143 f. Dissertação (Mestrado em Ciência da Informação) – Programa de pós-graduação em Ciência da Informação, Universidade Federal de Pernambuco, Recife, 2017. Disponível em: https://repositorio.ufpe.br/bitstream/123456789/25272/1/DISSERTA%C3%87%C3%83O%20Marcio%20Aercio%20Silva%20Bandim.pdf. Acesso em: 16 out. 2021.

BANDIM, M. A. S.; CORRÊA, R. F. Indexação automática por atribuição de artigos científicos em português da área de ciência da informação. **Transinformação**, [*s. l.*], v. 31, 2019. DOI: 10.1590/2318-0889201931e180004 Acesso em: 16 out. 2021.

BALAID, A.; ROZAN, M. Z. A.; HIKMI, S. N.; MEMON, J.; Knowledge maps: A systematic literatura review and directions for future research. **International Journal of Information Managemente**, v. 36, n. 3, p. 451-475. 2016. Disponível em: https://www.sciencedirect.com/science/article/pii/S0268401216000098?casa_token=PCBRSCabjqIAAAAA:M2cpQzCQNskh72ovxw3yXcm8DCNaIyShTkqA0htKe-BEetgKymxgGQHEsltMvhuRFHKA16Wc7w. Acesso em: 16 out. 2021.

BARKER, K.; CORNACCHIA, N. Using noun phrase heads to extract document keyphrases. *In:* CONFERENCE OF THE CANADIAN SOCIETY FOR COMPUTATIONAL STUDIES OF INTELLIGENCE**,** 1., 2000. Springer, Berlin, Heidelberg, p. 40-52, 2000. Disponível em: https://link.springer.com/chapter/10.1007/3-540-45486-1_4. Acesso em: 16 out. 2021.

BEZERRA, C. A.; GUIMARÃES, A. J. R. Mineração de texto aplicada às publicações científicas sobre gestão do conhecimento no período de 2003 a 2012. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 19, n. 2, p. 131-146, abr./jun. 2014. Disponível em: http://www.scielo.br/pdf/pci/v19n2/10.pdf. Acesso em: 16 out. 2021.

BIEBRICHER, P. *et al*. The automatic indexing system air/phys-from research to applications. *In:* **Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval**. 1988. p. 333-342. Disponível em: https://dl.acm.org/doi/abs/10.1145/62437.62470. Acesso em: 16 out. 2021.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **The Journal of Machine Learning Research**, v. 3, p. 993-1022, 2003. Disponível em: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?TB_iframe=true&width=370.8&height=658.8. Acesso em: 16 out. 2021.

BOLLEN, J. SOMPEL, H. V. HAGBERG, A. CHUTE, R. The principal component analysis of 39 scientific impact measures. **PLoS One**, São Francisco, v. 4, n.6, p.1-11, jun. 2009. Disponível em: https://doi.org/10.1371/journal.pone.0006022. Acesso em: 16 out. 2021.

BOOKSTEIN, A. *et al*. Probabilistic models for automatic indexing. **J. Am. Soc. Inf. Sci**., v. 25, n. 5, p. 312-316, 1974. Disponível em: http://qwone.com/~jason/papers/other/bookstein-indexing-74.pdf. Acesso em: 16 out. 2021.

BORGES, G. S. B.; LIMA, G. A. Desenvolvimento de softwares de indexação automática: breve avaliação dos principais critérios. **Informação & Tecnologia (ITEC)**, Marília/João Pessoa, v. 2, n.2, p. 49-70, jul./dez., 2015. Disponível em: https://periodicos.ufpb.br/ojs/index.php/itec/article/view/33926/17500. Acesso em: 16 out. 2021.

BORGMAN, C. L. Social aspects of digital libraries. *In:* WORKSHOP ON SOCIAL ASPECTS OF DIGITAL LIBRARIES, 1996, Los Angeles. **Final Report…** Los Angeles: UCLA/NSF, 1996.

BRAGA, F. R. Extração semiautomática de taxonomia para domínios especializados usando técnicas de mineração de textos. **Ciência da Informação**, Brasília, v.45, n.3, p.175-186, set./dez. 2016. Disponível em: http://revista.ibict.br/ciinf/article/view/4056. Acesso em: 16 out. 2021.

BRUNELLI, R.; MICH, O.; MODENA, C. M. A survey on the automatic indexing of video data. **Journal of Visual Communication and Image Representation**, [*s. l.*], v. 10, n. 2, p. 78-112, 1999. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1047320397904041. Acesso em: 16 out. 2021.

BUFREM, L. S.; SILVA, F. M. e; SOBRAL, N. V. Análise das influências intelectuais na produção científica da área de Ciência da Informação: um estudo sobre os bolsistas de produtividade em pesquisa (PQ-CNPq). **Em Questão**, Porto Alegre, v. 23, p.115-141. jan. 2017. Disponível em: http://seer.ufrgs.br/index.php/EmQuestao/article/view/68087. Acesso em: 16 out. 2021.

CHAUMIER, J. Indexação: conceito, etapas e instrumentos. **Revista Brasileira de Biblioteconomia e Documentação,** São Paulo, v. 21, n. 1, p. 63-79, 1988.

CHEBIL, W. *et al*. Indexation automatique de documents en santé: évaluation et analyse de sources d'erreurs. **IRBM**, v. 33, n. 5-6, p. 316-329, 2012. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1959031812001170. Acesso em: 16 out. 2021.

CHEN, M. *et al*. A practical system of keyphrase extraction for web pages. *In*: **Proceedings of the 14th ACM international conference on Information and knowledge management**. Bremen, p. 277-278. 2005. Disponível em: https://dl.acm.org/doi/10.1145/1099554.1099625. Acesso em: 16 out. 2021.

CHEN, X. *et al*. Topics and trends in artificial intelligence assisted human brain research. **PloS one**, [*s. l.*], v. 15, n. 4, p. 0231192, 2020. Disponível em: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231192. Acesso em: 16 out. 2021.

CHEN, X.; XIE, H. A structural topic modeling-based bibliometric study of sentiment analysis literature. **Cognitive Computation**, [*s. l.*], v. 12, n. 6, p. 1097-1129, 2020.

Disponível em: https://link.springer.com/article/10.1007/s12559-020-09745-1. Acesso em: 16 out. 2021.

CHIEN, L. F. PAT-tree-based keyword extraction for Chinese information retrieval. *In:* **Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval**. 1997. p. 50-58. Disponível em: https://dl.acm.org/doi/abs/10.1145/258525.258534. Acesso em: 16 out. 2021.

CHO, S. M.; PARK, C.; SONG, M. The evolution of social health research topics: A data-driven analysis. **Social Science & Medicine**, v. 265, p. 1-10, 2020. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0277953620305189. Acesso em: 16 out. 2021.

CHOI, S.; SEO, J. Y. An exploratory study of the research on caregiver depression: using bibliometrics and LDA topic modeling. **Issues in mental health nursing**, v. 41, n. 7, p. 592-601, 2020. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/01612840.2019.1705944. Acesso em: 16 out. 2021.

CHUMACHENKO, A. V.; KREMINSKYI, B. G.; MOSENKIS, I. L.; YAKIMENKO, A. I. Dynamics of topic formation and quantitative analysis of hot trends in physical science. **Scientometrics**, [*s. l.*], v. 125, n. 1, p. 739-753, 2020. Disponível em: https://link.springer.com/article/10.1007/s11192-020-03610-6. Acesso em: 16 out. 2021.

CORRÊA, R. F.; LAPA, R. C. Panorama de estudos sobre indexação automática no âmbito da ciência da informação no Brasil (1973-2012). **Ciência da Informação**, Brasília, v. 42, n. 2, p.255-273, 2013.

DAUD, A. *et al*. Finding rising stars through hot topics detection. **Future Generation Computer Systems,** [*s. l.*], v. 115, p. 798-813, 2021. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167739X20329903?casa_token=toZgkN0EYPcAAAAA:D3Nu8G94BOKfIz_NbX3JNA0qLWk9A1SvexGAYSFDG8XWkcIfbnWVlDjywp09rl7l3YvNsE5CVg. Acesso em: 16 out. 2021.

DE WINTER, J. C.; ZADPOOR, A. A.; DODOU, D. The expansion of Google Scholar versus Web of Science: a longitudinal study. **Scientometrics**, [*s. l.*], v. 98, n. 2, p. 1547-1565, fev. 2014. Disponível em: https://link.springer.com/article/10.1007%2Fs11192-013-1089-2. Acesso em: 16 out. 2021.

DING, Y.; CHOWDHURY, G. G.; FOO, S. Mapping the intellectual structure of information retrieval studies: an author co-citation analysis, 1987–1997. **Journal of Information Science**, v. 25, n. 1, p. 67-78, 1999. Disponível em: https://journals.sagepub.com/doi/10.1177/016555159902500107. Acesso em: 16 out. 2021.

DING, Y.; CHOWDHURY, G. G.; FOO, S. Bibliometric cartography of information retrieval research by using co-word analysis. **Information processing & management**, [*s. l.*], v. 37, n. 6, p. 817-842, 2001.

ERCAN, G.; CICEKLI, I. Using lexical chains for keyword extraction. **Information Processing & Management**, [*s. l.*], v. 43, n. 6, p. 1705-1714, 2007.

FARO, A.; GIORDANO, D.; SPAMPINATO, C. Combining literature text mining with microarray data: advances for system biology modeling. **Briefings in Bioinformatics**, Oxford, v.13, n.1, p. 61-82, 2011. Disponível em: https://academic.oup.com/bib/article/13/1/61/219461. Acesso em: 16 out. 2021.

FELDMAN, R.; DAGAN, I. Knowledge discovery in textual databases (KDT). *In:* **Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)**, Montreal, p. 20-21. 1995. Disponível em: https://www.aaai.org/Papers/KDD/1995/KDD95-012.pdf. Acesso em: 16 out. 2021.

FEOFILOFF, P.; KOHAYAKAWA, Y.; WAKABAYASHI, Y. **Uma introdução sucinta à teoria dos grafos**. São Paulo: UME/USP, 2011. 61 p.

FERNEDA, E. **Recuperação de informação:** Análise sobre a contribuição da Ciência da Computação para Ciência da Informação. 2003. 147p. Tese (Doutorado em Ciência da Informação) – Curso de Ciências da Comunicação, Departamento de Biblioteconomia e Documentação, Universidade de São Paulo, São Paulo 2003.

FERREIRA, M. da S. **A representação da memória científica da Ciência da Informação brasileira:** um estudo com as palavras-chave do ENANCIB. 2012. Dissertação (Mestrado em Ciência da Informação) – Programa de Pós-graduação em Ciência da Informação, Universidade Federal de Pernambuco, Recife, 2012.

FERREIRA, M. H. W.; CORRÊA, R. F. Estudo métrico sobre biblioteca digital: uso do software iramuteq. *In*: Encontro Nacional de Pesquisa em Ciência da Informação, 19., 2018, Londrina. **Anais do XIX ENANCIB**. Londrina: UEL, 2018. p. 4437-4454. Disponível em: http://hdl.handle.net/20.500.11959/brapci/102876. Acesso em: 16 out. 2021.

FIGUEROLA, C. G.; MARCO, F. J. G.; PINTO, M. Mapping the evolution of library and information science (1978–2014) using topic modeling on LISA. **Scientometrics**, v. 112, n. 3, p. 1507-1535, 2017. Disponível em: https://link.springer.com/article/10.1007/s11192-017-2432-9. Acesso em: 11 abr. 2021.

FINKELSTEIN, L. et al. Placing search in context: The concept revisited. *In:* **Proceedings of the 10th international conference on World Wide Web**. 2001. p. 406-414. Disponível em: https://dl.acm.org/doi/abs/10.1145/371920.372094. Acesso em: 16 out. 2021.

FRANCO, N. M. G.; FARIA, L. I. L. Colaboração científica intraorganizacional: análise de redes por coocorrência de palavras-chave. **Em Questão**, [*s. l.*], v. 25, n. 1, p. 87-110, 2019.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: An overview. **AI magazine**, v. 13, n. 3, p. 57-57, 1992. Disponível em: https://www.aaai.org/ojs/index.php/aimagazine/article/view/1011. Acesso em: 16 out. 2021.

FREITAS, C. M. D. S. *et al*. Introdução à visualização de informações. **Revista de informática teórica e aplicada,** Porto Alegre. v. 8, n. 2, p. 143-158, 2001. Disponível em: https://www.lume.ufrgs.br/handle/10183/19398. Acesso em: 16 out. 2021.

GARCÍA-MARCO, F. J.; FIGUEROLA, C. G.; PINTO, M. Análisis de la evolución temática de la investigación sobre Información y Documentación en español en la base de datos LISA mediante modelado temático (1978-2019). **Profesional de la Información**, Barcelona, v. 29, n. 4, 2020. Disponível em: https://recyt.fecyt.es/index.php/EPI/article/view/77144. Acesso em: 16 out. 2021.

GARFIELD, E. Historiographs, librarianship, and the history of science. *In:* CONRAD, H. R. (ed.). **Toward a theory of librarianship: papers in honor of Jesse Hauk Shera**. Philadelphia: Scarecrow Press, 1973.

GARVEY, W. D.; GRIFFITH, B. C. Scientific communication in social system. **Science**, Washington, v. 157, p.1011-1016, set. 1967. Disponível em: http://science.sciencemag.org/content/157/3792/1011. Acesso em: 16 out. 2021.

GIL, A. C. **Métodos e técnicas de pesquisa social**. São Paulo: Atlas, 2009.

GIL LEIVA, I. **Manual de indización: teoría y práctica**. Gijón: Ediciones Trea, 2008. 429 p.

GIL LEIVA, I. **La automatización de la indización, propuesta teórico-metodológica**: aplicación al área de Biblioteconomía y Documentación.1997. 268f. Tese (Doutorado em Filosofia e Letras) – Universidad de Murcia, Murcia, España, 1997. Disponível em: https://www.tesisenred.net/handle/10803/10917;jsessionid=5F13B4C0D6CED9D4F3E2DD7035865AE8#page=1. Acesso em: 16 out. 2021.

GIL LEIVA, I. SISA—Automatic Indexing System for Scientific Articles: Experiments with Location Heuristics Rules Versus TF-IDF Rules. **Knowledge Organization**, [*s. l.*], v. 44, n. 3, p. 139-162, 2017. Disponível em: https://webs.um.es/isgil/resources/SISA%20Automatic%20indexing%20Gil-Leiva2017.pdf. Acesso em: 16 out. 2021.

GOUVEIA, F. C. Altmetria: métricas de produção científica para além das citações Altmetrics: scientific production metrics beyond citations. **Liinc em revista**, [*s. l.*], v. 9, n. 1, 2013. p. 214-227, maio. 2013. Disponível em: http://revista.ibict.br/liinc/article/view/3434/3004. Acesso em: 16 out. 2021.

HASAN, K. S.; NG, V. Automatic keyphrase extraction: A survey of the state of the art. *In:* **Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics** (Volume 1: Long Papers). 2014. p. 1262-1273. Disponível em: https://www.aclweb.org/anthology/P14-1119.pdf. Acesso em: 16 out. 2021.

HEERSMINK, R.; HOVEN, J. V. D.; ECK, N. J. V.; BERG, J. V. D. Bibliometric mapping of computer and information ethics. **Ethics and information technology**, [*s. l.*], v. 13, n. 3, p.

241, 2011. Disponível em: https://link.springer.com/content/pdf/10.1007/s10676-011-9273-7.pdf. Acesso em: 16 out. 2021.

HEO, G. E. et al. Analyzing the field of bioinformatics with the multi-faceted topic modeling technique. **BMC bioinformatics**, [*s. l.*], v. 18, n. 7, p. 45-57, 2017. Disponível em: https://link.springer.com/article/10.1186/s12859-017-1640-x. Acesso em: 16 out. 2021.

HERSH, W. R.; GREENES, R. A. SAPHIRE, A. An information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. **Computers and Biomedical Research**, [*s. l.*], v. 23, n. 5, p. 410-425, 1990. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/0010480990900317. Acesso em: 16 out. 2021.

HU, K. *et al.* A domain keyword analysis approach extending Term Frequency-Keyword Active Index with Google Word2Vec model. **Scientometrics**, v. 114, n. 3, p. 1031-1068, 2018. Disponível em: https://link.springer.com/article/10.1007/s11192-017-2574-9. Acesso em: 16 out. 2021.

HUANG, C. *et al*. Keyphrase extraction using semantic networks structure analysis. *In:* Sixth International Conference on Data Mining (ICDM'06). **IEEE**, p. 275-284. 2006. Disponível em: https://ieeexplore.ieee.org/document/4053055. Acesso em: 16 out. 2021.

HULTH, A. Improved automatic keyword extraction given more linguistic knowledge. *In*: **Proceedings of the 2003 conference on Empirical methods in natural language processing**. 2003. p. 216-223. Disponível em: https://www.aclweb.org/anthology/W03-1028.pdf. Acesso em: 16 out. 2021.

HUMPHREY, S. M.; MILLER, N. E. Knowledge-based indexing of the medical literature: The Indexing Aid Project. **Journal of the American Society for Information Science**, [*s. l.*], v. 38, n. 3, p. 184-196, 1987. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/(SICI)1097-4571(198705)38:3%3C184::AID-ASI7%3E3.0.CO;2-F. Acesso em: 16 out. 2021.

IRVING, H. B. Computer-assisted indexing training and electronic text conversion at NAL. **Knowledge Organization**, [*s. l.*], v. 24, n. 1, p. 4-7, 1997. Disponível em: https://www.nomos-elibrary.de/10.5771/0943-7444-1997-1-4.pdf. Acesso em: 16 out. 2021.

JEONG, D. H.; SONG, M. Time gap analysis by the topic model-based temporal technique. **Journal of Informetrics**, v. 8, n. 3, p. 776-790, 2014. Disponível em: https://www.sciencedirect.com/science/article/pii/S1751157714000650?casa_token=jH5p1oQJcmkAAAAA:5SEE4ywqZ91-QO2MGFrMZOSmOQS4NErxzztR0NuL2ImcltCRS1ajpu1mP6WYf1vOE0J_5VYtMQ. Acesso em: 16 out. 2021.

JIANG, H. C.; QIANG, M. S.; LIN, P. Finding academic concerns of the Three Gorges Project based on a topic modeling approach. **Ecological indicators**, [*s. l.*], v. 60, p. 693-701,

2016. Disponível em: https://www.sciencedirect.com/science/article/pii/S1470160X15004288?casa_token=O-5YW-J2tG4AAAAA:4grcG-GWGaBe3RXG5gA14vLz2hkJBl9OwMsd2if001rcNGLabSjeFyUXLIyPIEnJL4oKNtKnpw. Acesso em: 16 out. 2021.

JONES, K. S. Automatic Indexing. **Jornal of Documetation**, Bingley, v. 30, p. 393-432, 1974. Disponível em: https://www.emeraldinsight.com/doi/abs/10.1108/eb026588. Acesso em: 16 out. 2021.

KELLEHER, D.; LUZ, S. Automatic hypertext keyphrase detection. *In*: **IJCAI**. Dublin, Irlanda. p. 1608-1609. 2005. Disponível em: https://pdfs.semanticscholar.org/2b47/3176be67cef1daa0379fa1616d5b7f248557.pdf. Acesso em: 16 out. 2021.

KIM, S.; PARK, H.; LEE, J. Word2vec-based latent semantic analysis (W2V-LSA) for topic modeling: A study on blockchain technology trend analysis. **Expert Systems with Applications**, v. 152, p. 113401, 2020. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417420302256. Acesso em: 16 out. 2021.

KLINGBIEL, P. H. Machine-aided indexing of technical literature. **Information Storage and Retrieval**, v. 9, n. 2, p. 79-84, 1973. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/002002717390020X. Acesso em: 16 out. 2021.

KOBASHI, N. Y.; DÍAZ, F.; SANTANA, S. Cartografia temática e de colaboração em organização do conhecimento no brasil (2000-2010). **Ciência da Informação**, v. 43, n. 1, 2014. DOI: 10.18225/ci.inf..v43i1.1417. Acesso em: 16 out. 2021.

KOBASHI, N. Y.; SANTOS, R. N. M. Institucionalização da pesquisa científica no Brasil: cartografia temática e de redes sociais por meio de técnicas bibliométricas. **Transinformação**, Campinas, v. 18, n. 1, p. 27-36, jan./abr., 2006. Disponível em: https://www.scielo.br/scielo.php?pid=S010378620060001000003&script=sci_abstract&tlng=pt. Acesso em: 16 out. 2021.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. ed. Brasília: Briquet de Lemos Livros, 2004. 452p.

LAPA, R. C. **Indexação automática no Brasil no âmbito da Ciência da Informação (1973-2012)**. 2014. 287f. Dissertação (Mestrado em Ciência da Informação) – Programa de Pós-graduação em Ciência da Informação, Universidade Federal de Pernambuco, Recife, 2014.

LAPA, R. C.; CORRÊA, R. F. Indexação automática no âmbito da ciência da informação no brasil. **Informação & Tecnologia**, [*s. l.*], v. 1, n. 2, p. 59-76, 2014. Disponível em: http://hdl.handle.net/20.500.11959/brapci/41624. Acesso em: 16 out. 2021.

LE COADIC, Y. F. **A Ciência da Informação.** Brasília: Briquet de Lemos, 2004.

LEITE, M. C. da. S. R.; ANDRADE, F. R. B. A. Nem sempre os opostos se atraem: o currículo do ensino médio e a práxis filosófica pós LDB nº 9.394/1996. **SAJEBTT**, Rio Branco, v. 7, n. 3, p. 213-235, 2020.

LIMA, P. G. S. G. **Aplicação do método de Louvain para extrair informações de grandes conjuntos de alertas de intrusão**. 2017. 73f. Monografia (Graduação em Ciência da Computação) – Bacharelado em Ciência da Computação, Universidade Estadual de Londrina, Londrina, 2017.

LIMA, T. V. **Professores de Matemática da Rede Estadual em Goiânia**: TDCI em perspectiva. 2017. 203f. Dissertação (Mestrado em Educação em Ciências e Matemática) – Programa de Mestrado em Educação em Ciências e Matemática, Universidade Federal de Goiás, Goiás, 2017. Disponível em: https://repositorio.bc.ufg.br/tede/handle/tede/7925. Acesso em: 16 out. 2021.

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. Disponível em: https://www.scielo.br/scielo.php?pid=S1413-99362009000100010&script=sci_arttext&tlng=pt. Acesso em: 16 out. 2021.

LIN, J. R. *et al*. Understanding On-Site Inspection of Construction Projects Based on Keyword Extraction and Topic Modeling. **IEEE Access**, v. 8, p. 198503-198517, 2020. Disponível em: https://ieeexplore.ieee.org/abstract/document/9246545. Acesso em: 16 out. 2021.

LOPES, I. L. Estratégia de busca na recuperação da informação: revisão da literatura. **Ciência da Informação**, v. 31, n. 2, p. 60-71, 2002. Disponível em: https://www.scielo.br/scielo.php?pid=S0100-19652002000200007&script=sci_arttext&tlng=pt. Acesso em: 16 out. 2021.

MAAREK, Y. S.; BERRY, D. M.; KAISER, G. E. An information retrieval approach for automatically constructing software libraries. **IEE Transactions on Software Engineering**, [*s. l.*], v. 17, n. 8, p. 800-813, 1991. Disponível em: https://academiccommons.columbia.edu/doi/10.7916/D81C251D. Acesso em: 10 dez. 2020.

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. Disponível em: http://www.brapci.inf.br/index.php/res/download/55796. Acesso em: 26 ago. 2019.

MARCELO, J. F.; HAYASHI, M. C. P. I. Estudo Bibliométrico sobre a Produção Científica no Campo da Sociologia da Ciência. **Informação e Informação**, Londrina, v. 18, n. 3, p. 138-153, set./dez. 2013. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/8413. Acesso em: 16 out. 2021.

MARCONI, M. de A.; LAKATOS, E. M. **Metodologia do Trabalho Científico**: procedimentos básicos, pesquisa bibliográfica, projeto e relatório, publicações e trabalhos científicos. São Paulo: Atlas, 2013.

MARCONDES, C. H.; COSTA, L. C.; MARTINS, S. C. Descoberta de conhecimento em artigos digitais em ciências biomédicas. **Informação & Informação**, Londrina, v. 21, n. 2, p. 170-216, 2016. DOI: 10.5433/1981-8920.2016v21n2p170. Acesso em: 16 out. 2021.

MARON, M. E. Automatic indexing: an experimental inquiry. **Journal of the ACM JACM**, [*s. l.*], v. 8, n. 3, p. 404-417, 1961. Disponível em: https://sci2s.ugr.es/keel/pdf/algorithm/articulo/Maron1961.pdf. Acesso em: 16 out. 2021.

MARTÍNEZ, G. E. R. **Análisis comparativo entre los algoritmos de Kamada Kawai & Kruskal para la creación, estructura y optimización de portafolios de inversión a través de árboles de mínima expansión**. 2014. 56f. Dissertação (Mestrado em Finanças) – Instituto Tecnológico e de Estudos Superiores de Monterrey, EGADE Escola de Negócios, 2014.

MEDELYAN, O. **Human-competitive automatic topic indexing**. Nova Zelândia, 2009. 214f. Tese (Doutorado) - Department of Computer Science, The University of Waikato. Nova Zelândia, 2009.

MEDELYAN, O.; WITTEN, I. H. Thesaurus based automatic keyphrase indexing. *In:* **Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries**. p. 296-297. 2006. Disponível em: https://www.cs.waikato.ac.nz/~ihw/papers/06-OM-IHW-Thesaurus-auto-keyphrase.pdf. Acesso em: 16 out. 2021.

MONTEJO R. A. Proyecto de indexado automático para documentos en el campo de la física de altas energías. **Procesamiento del lenguaje natural**, n. 27, p. 295-296, set. 2001. Disponível em: http://rua.ua.es/dspace/handle/10045/1824. Acesso em: 16 out. 2021.

MUGNAINI, R.; JANNUZZI, P. de M.; QUONIAM, L. Indicadores bibliométricos da produção científica brasileira: uma análise a partir da base Pascal. **Ciência da Informação**, Brasília, v. 33, n. 2, p. 123-131, ago. 2004. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652004000200013&lng=en&nrm=iso. Acesso em: 16 out. 2021.

NADZAR, N. M. A. M.; BAKRI, A.; IBRAHIM, R. A Bibliometric Mapping of Malaysian Publication using Co-Word Analysis. **International Journal of Advances in Soft Computing and its Applications**, Jordan, v. 9, n. 3, 2017. ISSN 2074-8523. Disponível em: https://www.semanticscholar.org/paper/A-bibliometric-mapping-of-malaysian-publication-Nadzar-Bakri/d9dc91d36cdc9a3a2cc87ca9cca572111bcb7751?p2df. Acesso em: 16 out. 2021.

NAGARKAR, S. P.; KUMBHAR, R.; Text mining: An analysis of research published under the subject category 'Information Science Library Science' in Web of Science Database during 1999-2013. **Library Review**, [*s. l.*], v. 64, n. 3, 2015.

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

NÉVÉOL, A., SHOOSHAN, S. E., HUMPHREY, S. M., MORK, J. G., ARONSON, A. R. A recent advance in the automatic indexing of the biomedical literature. **Journal of biomedical informatics**, Amsterdã, v. 42, n. 5, p. 814-823, 2009. Disponível em: https://www.sciencedirect.com/science/article/pii/S1532046408001561. Acesso em: 05 mar. 2020.

NGUYEN, T. D.; KAN, M. Y. Keyphrase extraction in scientific publications. *In:* **International conference on Asian digital libraries**. Springer, Berlin, Heidelberg, 2007. p. 317-326. Disponível em: https://link.springer.com/chapter/10.1007/978-3-540-77094-7_41. Acesso em: 16 out. 2021.

NI, P.; LI, Y.; CHANG, V. Research on Text Classification Based on Automatically Extracted Keywords. **International Journal of Enterprise Information Systems (IJEIS)**, v. 16, n. 4, p. 1-16, 2020. Disponível em: https://www.igi-global.com/article/research-on-text-classification-based-on-automatically-extracted-keywords/265122. Acesso em: 26 jan. 2021.

NORONHA, D. P.; DE MELO MARICATO, J. Estudos métricos da informação: primeiras aproximações. **Encontros Bibli**, [*s. l.*], v. 13, n. 1, p. 116-128, 2008. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2008v13nesp1p116. Acesso em: 13 jan. 2021.

NOYONS, E.; VAN RAAN, A. Bibliometric cartography of scientific and technological developments of an R & D field: The case of optomechatronics. **Scientometrics**, v. 30, n. 1, p. 157-173, 1994. Disponível em: https://link.springer.com/article/10.1007/BF02017220. Acesso em: 16 out. 2021.

OH, J.; LEE, B. G. A Technical Approach for Suggesting Research Directions in Telecommunications Policy. **KSII Transactions on Internet & Information Systems**, v. 8, n. 12, p. 4467-4488, 2014. Disponível em: http://apps-webofknowledge.ez10.periodicos.capes.gov.br/full_record.do?product=WOS&search_mode=GeneralSearch&qid=3&SID=8FoGlkM6MdUuWjH3VXh&page=1&doc=1&cacheurlFromRightClick=no. Acesso em: 16 out. 2021.

OHSAWA, Y.; BENSON, N. E.; YACHIDA, M. KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. *In:* **Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries**-ADL'98-. IEEE, 1998. p. 12-18. Disponível em: https://ieeexplore.ieee.org/abstract/document/670375. Acesso em: 16 out. 2021.

OKUBO, Y. **Bibliometric Indicators and analysis of research systems**: methods and examples. OECD - Science, Technology and Industry Working Papers 1997/1. Paris: OECD Publishing, 1997. Disponível em: http://www.oecd-ilibrary.org/docserver/download/208277770603.pdf?expires=1494282294&id=id&accname=guest&checksum=EDB026194486F73E2EAA5F615A82F127. Acesso em: 16 out. 2021.

OLIVEIRA, E. F. T. de; GRÁCIO, M. C. C. Indicadores bibliométricos em ciência da informação: análise dos pesquisadores mais produtivos no tema estudos métricos na base Scopus. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 16, n. 4, p. 16-28, dez. 2011. Disponível em: https://www.scielo.br/scielo.php?pid=S1413-99362011000400003&script=sci_arttext&tlng=pt. Acesso em: 16 out. 2021.

ONAN, A.; KORUKOĞLU, S.; BULUT, H. Ensemble of keyword extraction methods and classifiers in text classification. **Expert Systems with Applications**, v. 57, p. 232-247, 2016. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0957417416301464. Acesso em: 16 out. 2021.

PANSANI JUNIOR, E. A.; FERNEDA, E. Ontologias no Processo de Indexação Automática de Documentos Textuais. **Anais do Encontro Nacional em Ciência da Informação.** Salvador, v. 17, p. 1-20, set./dez. 2016. Disponível em: http://repositorios.questoesemrede.uff.br/repositorios/handle/123456789/3616. Acesso em: 16 out. 2021.

PÉREZ-GUADARRAMAS, Y. *et al.* Analysis of OWA operators for automatic keyphrase extraction in a semantic context. **Intelligent Data Analysis**, v. 24, n. S1, p. 43-62, 2020. Disponível em: https://content.iospress.com/articles/intelligent-data-analysis/ida200008. Acesso em: 16 out. 2021.

PÉREZ-GUADARRAMAS, Y. *et al*. A Fuzzy Approach to Improve an Unsupervised Automatic Keyphrase Extraction Process. *In*: **2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)**. IEEE, 2018. p. 1-6. Disponível em: https://ieeexplore.ieee.org/abstract/document/8491487. Acesso em: 26 jan. 2021.

PINTO, A. L.; MATIAS, M. Indicadores Científicos e as Universidades Brasileiras. **Informação e Informação**, Londrina, v. 16, n. 3, p. 1-18, jan./jun., 2011. Disponível em: https://brapci.inf.br/index.php/res/download/44291. Acesso em: 16 out. 2021.

PINTO, D. M.; CORDEIRO, F. L.; TAKEMURA, C. M.; SOLANO, V. de O. Cartografia Temática da Produção Ténico-Científica da Embrapa Destinada à Agricultura Familiar. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 13, n. esp., p. 392-410, 2017. Disponível em: https://brapci.inf.br/index.php/res/download/40209. Acesso em: 16 out. 2021.

PULGARÍN, A.; GIL LEIVA, I. Bibliometric analysis of the automatic indexing literature: 1956 – 2000. **Information Processing and Management**, Amsterdã, v. 40, n. 2, p. 365-377, nov. 2004. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0306457302001012. Acesso em: 16 out. 2021.

RANAEI, S. *et al*. Evaluating technological emergence using text analytics: two case technologies and three approaches. **Scientometrics**, v. 122, n. 1, p. 215-247, 2020. Disponível em: https://link.springer.com/article/10.1007/s11192-019-03275-w. Acesso em: 16 out. 2021.

RANAEI, S.; SUOMINEN, A. Using machine learning approaches to identify emergence: Case of vehicle related patent data. *In:* **2017 Portland International Conference on Management of Engineering and Technology (PICMET)**. IEEE, 2017. p. 1-8. Disponível em: https://ieeexplore.ieee.org/abstract/document/8125290?casa_token=8wFjxeldAksAAAAA:NnstuH8uWsN17Antuqp032BSyqQyqDwdtGeDYc5bnyYUQ49te0GcE2X4eppU7FgXkk92o_pBEw. Acesso em: 16 out. 2021.

REZENDE, S. O. (org.) **Sistemas Inteligentes**: fundamentos e aplicações. São Paulo: Manole: 2005.

ROBREDO, J. A indexação automática de textos: o presente já entrou no futuro. *In:* Machado, U. O. **Estudos Avançados em Biblioteconomia e Ciência da Informação**. Brasília: ABDF, 1983. v. 1, n. 1, 1982, p. 235-274.

ROZA, R. H. Ciência da informação, tecnologia e sociedade. **Biblos**, [S. l.], v. 32, n. 2, p. 177–190, 2018. Disponível em: https://www.seer.furg.br/biblos/article/view/7546. Acesso em: 16 out. 2021.

SANTOS, C. A. C. M. dos. Organização e representação do conhecimento: bibliometria temática em artigos de periódicos brasileiros. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 11, n. esp., p. 640-653, 2015. Disponível em: https://rbbd.febab.org.br/rbbd/article/view/494. Acesso em: 16 out. 2021.

SALISBURY, L.; SMITH, J. J. Building the AgNIC Resource Database Using Semi-Automatic Indexing of Material. **Journal of Agricultural & Food Information**, [*s. l.*], v. 15, n. 3, p. 159-176, 2014. Disponível em: https://www.tandfonline.com/doi/abs/10.1080/10496505.2014.919805. Acesso em: 16 out. 2021.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. **Communications of the ACM**, New York, v. 18, n. 11, p. 613-620, mar. 1975. Disponível em: https://dl.acm.org/doi/abs/10.1145/361219.361220. Acesso em: 16 out. 2021.

SALTON, G.; YANG, C. S. On the specification of term values in automatic indexing. **Jornal of Documentation**, Ithaca, v. 29, n. 4, p. 351-372, dez. 1973. Disponível em: https://ecommons.cornell.edu/handle/1813/6016. Acesso em: 16 out. 2021.

SÉRGIO, M. C.; SILVA, T. do N. da; GONÇALVES, A. L. Descoberta de conhecimento a partir de informações não estruturadas por meio de técnicas de correlação e associação. **Em Questão**, Porto Alegre, v. 22, n. 2, p. 87-113, maio/ago. 2016. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/59514. Acesso em: 16 out. 2021.

SHAIKH, Z. A. Keyword detection techniques: a comprehensive study. Engineering, **Technology & Applied Science Research**, v. 8, n. 1, p. 2590-2594, 2018. Disponível em: http://www.etasr.com/index.php/ETASR/article/view/1813. Acesso em: 16 out. 2021.

SHIN, S. H. et al. Analyzing sustainability literature in maritime studies with text mining. **Sustainability**, v. 10, n. 10, p. 3522, 2018. Disponível em: https://www.mdpi.com/2071-1050/10/10/3522. Acesso em: 16 out. 2021.

SHIRMOHAMMADI, M.; HEDAYATI MEHDIABADI, A.; BEIGI, M.; MCLEAN, G. N. Mapping human resource development: Visualizing the past, bridging the gaps, and moving toward the future. **Human Resource Development Quarterly**, [*s. l.*], v. 28, n. 1, p. 1-28, 2020. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/hrdq.21415. Acesso em: 16 out. 2021.

SILVA, F. M. *et al*. Analysis of the communities of Brazilian researchers in the area of Philosophy: a study based on the juxtaposition between the data of the Lattes Platform and Web of Science (2007-2016). **Informacao & Sociedade-Estudos**, v. 28, n. 3, p. 245-262, 2018. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/41223/21828. Acesso em: 16 out. 2021.

SILVA, S. R. de B. **Sistemas de Indexação Automática por Atribuição**: uma análise comparativa. Dissertação (Mestrado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação, Universidade Federal de Pernambuco, Recife, 2020. 190f. Disponível em: https://repositorio.ufpe.br/bitstream/123456789/37626/1/DISSERTA%C3%87%C3%83O%20S%C3%A2mela%20Rouse%20de%20Brito%20Silva.pdf. Acesso em: 16 out. 2021.

SILVA, S. R. de B.; CORRÊA, R. F. Sistemas de Indexação Automática por Atribuição: uma análise comparativa. **Encontros Bibli**, Florianópolis, v. 25, p. 01-15, 2020. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2020.e70740/43771. Acesso em: 16 out. 2021.

SILVEIRA, M. A. A. da; CAREGNATO, S. E. Demarcações epistemológicas dos estudos de citação: concepção sociocultural das citações. **Perspectivas em Ciência da Informação**, [*s. l.]*, v. 23, p. 55-70, 2018. Disponível em: https://www.scielo.br/j/pci/a/FyhdsCYR5fK9FBsrkgYYz9v/?lang=pt&format=pdf. Acesso em: 16 out. 2021.

SIMÕES, M. G. M.; MACHADO, L. M. O.; SOUZA, R. R.; LOPES, A. T. Indexação automática e ontologias: identificação dos contributos convergentes na ciência da informação. **Ciência da Informação**, Distrito Federal, v. 46, n.1, p. 141-151, 2017. Disponível em: http://revista.ibict.br/ciinf/article/view/4020/3459. Acesso em: 16 out. 2021.

SINKKILÄ, R.; SUOMINEN, O.; HYVÖNEN, E. Automatic semantic subject indexing of web documents in highly inflected languages. *In*: **Extended Semantic Web Conference**. Springer, Berlin, Heidelberg, 2011. p. 215-229. Disponível em: https://link-springer-

com.ez10.periodicos.capes.gov.br/content/pdf/10.1007%2F978-3-642-21034-1_15.pdf. Acesso em: 16 out. 2021.

SIQUEIRA, J. C. Recursos linguísticos para análise de vocabulário controlado: o caso do SAUSP. **Biblionline**, João Pessoa, v. 7, n. 2, p. 52-62, 2011.

SMALL, H. Cocitation in the scientific literature: a new measure of the relationship between two documents. **Journal of the American Society for Information Science**, [*s. l.*], v. 24, p. 265-269. 1973. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.4630240406. Acesso: 16 out. 2021.

SOBRAL, N. V. **Pesquisadores em Doenças Tropicais Negligenciadas no Brasil**: produção científica e convergências com o plano nacional de saúde (2016 a 2019). 2019. 214 f. Tese de Doutorado (Doutorado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação, Universidade Federal da Bahia. Salvador, 2019.

SRINAVASAN, P. Text Mining: Generating Hypotheses From MEDLINE. **Journal of The American Society for Information Science and Technology**, [*s. l.*], v. 55, n. 5, p. 396-413, 2004. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.10389. Acesso em: 16 out. 2021.

STEVENS, M. E. **Indexação automática**: um estado da arte. Washington: National Bureau of Standards, 1965.

SONG, M.; HEO, G. E.; LEE, D. Identifying the landscape of Alzheimer's disease research with network and content analysis. **Scientometrics**, v. 102, n. 1, p. 905-927, 2015. Disponível em: https://link.springer.com/article/10.1007%2Fs11192-014-1372-x. Acesso em: 16 out. 2021.

SOUZA, R. R. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. 2005. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação da Universidade Federal de Minas Gerais, Belo Horizonte, 2005. Disponível em: http://livros01.livrosgratis.com.br/cp078465.pdf. Acesso em: 16 out. 2021.

TARAPANOFF, K. Referencial teórico: introdução. *In*: TARAPANOFF, K. (Org.). **Inteligência organizacional e competitiva**. Brasília: UnB, 2001. p.33-58.

TARGINO, M. G. A. Comunicação científica: uma revisão de seus elementos básicos. **Informação & Sociedade**: Estudos, João Pessoa, v. 10, n. 2, p. 37-85, 2000. Disponível em: http://basessibi.c3sl.ufpr.br/brapci/v/a/1182. Acesso em: 16 out. 2021.

TIMAKUM, T.; KIM, G.; SONG, M. A data-driven analysis of the knowledge structure of library science with full-text journal articles**. Journal of librarianship and information science**, v. 52, n. 2, p. 345-365, 2020. Disponível em: https://journals.sagepub.com/doi/full/10.1177/0961000618793977. Acesso em: 16 out. 2021.

TRUCOLO, C. C.; DIGIAMPIETRI, L. A. Análise de tendências da produção científica nacional na área de ciência da informação: estudo exploratório de mineração de textos. **AtoZ**: Novas Práticas em Informação e Conhecimento, Curitiba, v. 3, n. 2, p. 87-94, 2014.

Disponível em: https://revistas.ufpr.br/atoz/article/view/41341/25335. Acesso em: 16 out. 2021.

TRZESNIAK, P. Indicadores quantitativos: como obter, avaliar, criticar e aperfeiçoar. **Navus -** Revista de Gestão e Tecnologia, Florianópolis, v. 4, n. 2, p. 05-18, jul./dez. 2014. Disponível em: http://navus.sc.senac.br/index.php/navus/article/view/223. Acesso em: 16 out. 2021.

TSENG, Yuen-Hsien; LIN, Chi-Jen; LIN, Yu-I. Text mining techniques for patent analysis. **Information Processing & Management**, v. 43, n. 5, p. 1216-1247, 2007.

TURNEY, P. D. Learning algorithms for keyphrase extraction. **Information retrieval**, v. 2, n. 4, p.303-336, 2000. Disponível em: https://arxiv.org/ftp/cs/papers/0212/0212020.pdf. Acesso em: 16 out. 2021.

URBIZAGÁSTEGUI ALVARADO, R. A. A Lei de Lotka na bibliometria brasileira. **Ciência da Informação**, v.31, n.2, p. 14-20, maio/ago. 2002. Disponível em: http://www.scielo.br/pdf/ci/v31n2/12904.pdf. Acesso em: 16 out. 2021.

URBIZAGÁSTEGUI ALVARADO, R. A. Bibliometria no Brasil. **Ciência da Informação**, Brasília, v.13, n.2, p. 91-105, jul./dez. 1984. Disponível em: http://www.brapci.inf.br/index.php/res/download/53184. Acesso em: 16 out. 2021.

URBIZAGÁSTEGUI ALVARADO, R. A. Cientometria como um campo científico. **Informação & Sociedade**: Estudos, João Pessoa, v.20, n.3, p.41-62, set./dez. 2010.

VAN ECK, N. J.; WALTMAN, L. Software survey: VOSviewer, a computer program for bibliometric mapping. **Scientometrics**, v. 84, n. 2, p. 523-538, 2009. Disponível em: https://link.springer.com/content/pdf/10.1007/s11192-009-0146-3.pdf. Acesso em: 16 out. 2021.

VAN ECK, N. J.; WALTMAN, L.; DEKKER, R.; BERG, J. V. D. A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. **Journal of the American Society for Information Science and Technology**, v. 61, n. 12, p. 2405-2416, 2010. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21421. Acesso em: 16 out. 2021.

VAN ECK, N. J.; WALTMAN, L.; NOYONS, E. C. M.; BUTER, R. K. Automatic term identification for bibliometric mapping. **Scientometrics**, v. 82, n. 3, p. 581-596, 2010. Disponível em: https://akjournals.com/view/journals/11192/82/3/article-p581.xml. Acesso em: 16 out. 2021.

VANZ, S. A. S.; STUMPF, I. R. C. Procedimentos e Ferramentas Aplicados aos Estudos Bibliométricos. **Informação & Sociedade**: Estudos, João Pessoa, v.20, n.2, p. 67-75, maio/ago. 2010. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/4817. Acesso em: 16 out. 2021.

VIANA, A. S. **Temáticas das Teses dos Programas de Pós-Graduação em Ciência da Informação Nível Seis na CAPES**. 2016. Dissertação (Mestrado em Ciência da Informação) – Escola de Ciência da Informação da Universidade Federal de Minas Gerais, Belo Horizonte, 2016. Disponível em: https://repositorio.ufmg.br/handle/1843/BUBD-AE3JQF. Acesso em: 16 out. 2021.

VIEIRA, S. B. Indexação Automática e Manual: revisão de literatura. **Ciência da Informação**, Brasília, v. 17, n. 1, p. 43-57, jan./jun. 1988. Disponível em: http://revista.ibict.br/ciinf/article/view/298/298. Acesso em: 16 out. 2021.

VOGEL, M. J. M.; KOBASHI, N. Y. Tesauro funcional para organização de arquivos administrativos. Páginas A&B, **Arquivos e Bibliotecas (Portugal)**, n. 12, p. 48-62, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/127667. Acesso em: 16 out. 2021.

WITTEN, I. H.; FRANK, E.; HALL, M. **Data Mining**: Practical Machine Learning Tools and Techniques. 3. ed. New York: Morgan Kaupmann, 2011.

WITTEN, I. H. *et al*. Kea: Practical automated keyphrase extraction. *In:* Design and Usability of Digital Libraries: Case Studies in the Asia Pacific. **IGI global**, Hershey, 2005. p. 129-152. Disponível em: https://www.cs.waikato.ac.nz/~ml/publications/2005/chap_Witten-et-al_Windows.pdf. Acesso em: 16 out. 2021.

YEO, J. S.; JEONG, Y. Pathway toward market entry of perovskite solar cells: A detailed study on the research trends and collaboration networks through bibliometrics. **Energy Reports**, v. 6, p. 2075-2085, 2020. Disponível em: https://www.sciencedirect.com/science/article/pii/S235248472031249X. Acesso em: 16 out. 2021.

YU, C. T.; SALTON, G. Precision weighting—an effective automatic indexing method. **Journal of the ACM (JACM)**, v. 23, n. 1, p. 76-88, 1976. Disponível em: https://dl.acm.org/doi/abs/10.1145/321921.321930. Acesso em: 16 out. 2021.

ZHANG, Y. CAO, B. L.; WANG, Y. F.; PENG, T. Q.; WANG, X. H. When Public Health Research Meets Social Media: Knowledge Mapping From 2000 to 2018. **Journal of medical Internet research**, v. 22, n. 8, p. e17582, 2020. Disponível em: https://www.jmir.org/2020/8/e17582/. Acesso em: 02 fev. 2021.

APÊNDICE A – RESULTADOS DO MAUI DO *CORPUS* A

Medidas alcançadas	Precisão: 56,54% +/- 18,7% Revocação: 45,29% +/- 18,06% Medida – F: 50,29%
ARTIGO 01	conhecimento nas organizações transferência da informação gestão do conhecimento descarte avaliação acesso gestão
ARTIGO 02	comunicação científica divulgação científica ciência da informação notícias educação comunicação
ARTIGO 03	direito à informação recuperação da informação avaliação direito confiabilidade confiabilidade
ARTIGO 04	acesso livre direito autoral arquivos livros filosofia conceitos publicações arranjo direito acesso
ARTIGO 05	comunicação científica informação científica e tecnológica sumários educação estudantes cientistas acesso comunicação pesquisa
ARTIGO 06	ciência da computação ciência da informação interdisciplinaridade recuperação da informação sistemas de informação navegação

	computadores gestão pesquisa tesauros
ARTIGO 07	transferência da informação recursos de informação gestão do conhecimento inteligência competitiva gestão da informação competitividade papel dados gestão
ARTIGO 08	conceitos de informação estados da arte ciência da informação áreas do conhecimento homografia artes registros pesquisa termos conceitos
ARTIGO 09	conceitos de informação tomada de decisões disseminação seletiva da informação fluxo da informação ciência da informação inteligência competitiva gestão gestores universidades conceitos
ARTIGO 10	probabilidade e estatística pertinência levantamentos economia
ARTIGO 11	ciência da informação propriedade intelectual recuperação da informação educação classificação pesquisa dados
ARTIGO 12	acesso à informação arquivologia sistemas de informação áreas do conhecimento ciência da informação nomes próprios

	acesso biblioteconomia educação pesquisa
ARTIGO 13	profissionais de informação sistemas de informação formação profissional bibliotecas economia usuários papel acervos
ARTIGO 14	inclusão digital competência em informação gestão do conhecimento profissionais de informação gestão da informação necessidades de informação gestão acesso papel gestores
ARTIGO 15	formação profissional profissionais de informação mudança universidades pesquisa educação
ARTIGO 16	educação superior sociedade da informação formação profissional ciência da informação profissionais de informação mudança biblioteconomia educação
ARTIGO 17	controle bibliográfico tecnologias da informação e comunicação publicações eletrônicas propriedade intelectual direito à privacidade comunicação publicações relevância autoria complexidade
ARTIGO 18	sociologia categorias
ARTIGO 19	ambiente organizacional gestão do conhecimento

	gestão da informação inteligência competitiva avaliação bibliotecas gestão pesquisa confiabilidade confiabilidade
ARTIGO 20	ciência da informação gráficos publicações pesquisa dissertações e teses avaliação
ARTIGO 21	inteligência artificial áreas do conhecimento multidisciplinaridade sociologia autoria usuários dissertações e teses
ARTIGO 22	tecnologias da informação e comunicação processamento de informações ciência da informação hipertextos hipertexto formatos MARC serviços de empréstimo paradigmas pesquisa biblioteconomia
ARTIGO 23	relações entre conceitos disseminação da informação ciência da informação história
ARTIGO 24	ecologia validade buscas pesquisadores pesquisa
ARTIGO 25	ciência da informação agências de fomento programas de pós-graduação orçamento levantamentos pesquisadores pesquisa dados
ARTIGO 26	ciência da informação programas de computador

	subsídios semântica administração monitoramento gestão
ARTIGO 27	bases de dados monitoramento agências de fomento indicadores de C&T indicadores gestão relevância pesquisadores dados
ARTIGO 28	bibliotecas digitais informação governamental gestão da informação necessidades de informação governo eletrônico sistemas de informação inovação bibliotecas papel dissertações e teses
ARTIGO 29	políticas de informação governo eletrônico gestão da informação informação governamental gestão identificadores de objetos digitais acesso administração
ARTIGO 30	periódicos eletrônicos World Wide Web avaliação periódicos arquivos acesso
ARTIGO 31	regimes de informação sociedade da informação políticas de informação Internet serviços de empréstimo cenários semiótica conceitos economia termos
ARTIGO 32	linguagens documentárias buscas de informação

	revisões de literatura estratégias de busca bases de dados recuperação da informação planejamento discos compactos termos complexidade
ARTIGO 33	referências bibliográficas bibliotecas virtuais bibliotecas digitais tipos de documento desenvolvimento de coleções unidades de informação produtividade de autor artigos de periódico profissionais de informação planejamento
ARTIGO 34	ciências da saúde informação científica e tecnológica transferência da informação nutrição mudança acesso comunicação universidades
ARTIGO 35	linguagens de marcação ciência da informação intranetes HTML XML editores normas e protocolos editoras publicações Internet
ARTIGO 36	arquivos privados produtividade de autor lei de Lotka cartas periódicos medicina arquivos bibliometria autoria dados
ARTIGO 37	informação jurídica informação financeira indústria da informação informação para negócios

	tomada de decisões bases de dados guias classificação cenários empresas
ARTIGO 38	bibliotecas híbridas tecnologias da informação e comunicação ensino a distância usuários Internet papel comunicação universidades bibliotecas educação
ARTIGO 39	sistemas de recuperação da informação buscas de informação revisões de literatura estratégias de busca bases de dados recuperação da informação planejamento acesso conceitos usuários
ARTIGO 40	direito autoral mercado de trabalho recursos de informação ciência da informação profissionais de informação gestão da informação direito à privacidade direito gestores dados
ARTIGO 41	autoria individual literatura cinzenta artigos de periódico institutos de pesquisa ciência da informação índices índice h periódicos autoria dados
ARTIGO 42	bibliotecas virtuais informação para negócios gestão do conhecimento sistemas de informação

	bibliotecas gestão da informação inteligência competitiva profissionais de informação empresas usuários
ARTIGO 43	informação científica e tecnológica necessidades de informação empresas inovação pesquisa
ARTIGO 44	referências bibliográficas tecnologias da informação e comunicação gestão do conhecimento conhecimento nas organizações empresas competitividade papel gestão
ARTIGO 45	fluxo da informação ciência da informação Internet registros bibliometria similaridade informetria avaliação comunicação autoria
ARTIGO 46	métodos de pesquisa lei de Bradford teoria do caos ciência da informação bibliometria física periódicos conceitos pesquisa identificadores de objetos digitais
ARTIGO 47	periódicos científicos periódicos eletrônicos ciência da informação usabilidade usuários periódicos World Wide Web
ARTIGO 48	usos da informação economia da informação custos inovação

	empresas buscas levantamentos economia
ARTIGO 49	pesquisa e desenvolvimento ciência da informação descritores periódicos títulos de documentos programas de computador World Wide Web pesquisa biblioteconomia autoria
ARTIGO 50	cooperação formação profissional profissionais de informação bibliotecários conceitos papel educação
ARTIGO 51	acessibilidade categorias usuários bibliotecas
ARTIGO 52	projetos de pesquisa estudos de usuários ciência da informação normas e protocolos hipertextos hipertexto Internet papel educação pesquisa
ARTIGO 53	mecanismos de busca recuperação da informação estudos experimentais professores estudantes World Wide Web usuários buscas universidades
ARTIGO 54	sociedade da informação competência em informação paradigmas filosofia bibliotecários educação

	bibliotecas avaliação acesso conceitos
ARTIGO 55	mercado de trabalho profissionais de informação ciência da informação imagens empresas consultores gestores
ARTIGO 56	comunicação científica teorias na ciência da informação cientistas da informação organização do conhecimento história da ciência da informação informação científica e tecnológica ciência da informação normas e protocolos sociologia história
ARTIGO 57	recuperação da informação ciência da informação pesquisa avaliação
ARTIGO 58	ciência cognitiva processamento de informações recuperação da informação ciência da informação indexação computadores
ARTIGO 59	comunicação científica frente de pesquisa citações bibliográficas análise de citação estruturalismo cientometria bibliometria comunicação educação pesquisa
ARTIGO 60	agentes inteligentes estudos de caso processos de gestão tecnologias da informação e comunicação usos da informação aplicações de computador gestão da informação inteligência competitiva inovação

	Internet

APÊNDICE B – RESULTADOS DO MAUI DO *CORPUS* B

Medidas Alcançadas	Precisão: 51.31% +/- 13% Revocação: 85.09% +/- 14.46% Medida-F: 64.02%
ARTIGO 01	bibliotecas digitais funcionalidade bibliotecas universidades dissertações e teses
ARTIGO 02	políticas públicas projetos de pesquisa tecnologias da informação e comunicação bibliotecas digitais imagens educação pesquisa comunicação acesso bibliotecas
ARTIGO 03	cooperação entre bibliotecas dados científicos bibliotecas universitárias bibliotecas digitais periódicos eletrônicos desenvolvimento de coleções gestores de bibliotecas unidades de informação livros eletrônicos gestão da informação
ARTIGO 04	bibliotecas universitárias bibliotecas centrais preservação digital bibliotecas digitais análise qualitativa revisões de literatura entrevistas gestão buscas conceitos
ARTIGO 05	objetos digitais sistemas de recuperação da informação bibliotecas digitais disseminação da informação recuperação da informação ciência da informação mapas usuários bibliotecas

ARTIGO 06	documentos primários objetos digitais classificação automática teoria do conceito bibliotecas digitais ciência da informação recuperação da informação bibliotecários bibliometria professores
ARTIGO 07	brecha digital bibliotecas universitárias bibliotecas digitais cooperação questionários bibliotecários entrevistas estudantes indicadores de C&T indicadores
ARTIGO 08	bibliotecas digitais categorias bibliografias levantamentos bibliotecas avaliação pesquisa
ARTIGO 09	repositórios institucionais bibliotecas digitais economia da informação estudos de caso bibliotecas universitárias acesso livre coleções programas de computador inovação arquivos
ARTIGO 10	bibliotecas digitais direito autoral engenharia biomédica informação científica e tecnológica direito à informação tomada de decisões propriedade intelectual gestão da informação subsídios engenharias
ARTIGO 11	bibliotecas digitais ciências sociais aplicadas sociedade da informação

	acesso livre professores coleções livros digitalização acervos acesso
ARTIGO 12	bibliotecas digitais informação governamental gestão da informação necessidades de informação governo eletrônico sistemas de informação inovação bibliotecas papel dissertações e teses
ARTIGO 13	arquitetura de informação bibliotecas digitais disseminação da informação acesso bibliotecas usuários
ARTIGO 14	bibliotecas universitárias organização do conhecimento redes de bibliotecas tecnologias da informação e comunicação bibliotecas virtuais bases de dados bibliotecas digitais superposição estágios bibliotecários
ARTIGO 15	bibliotecas digitais estudos de caso redes de telecomunicações recuperação da informação usabilidade eficiência pesquisa acesso dissertações e teses buscas
ARTIGO 16	direito autoral arquitetura de informação desenvolvimento de coleções preservação digital bibliotecas digitais acesso à informação controle bibliográfico

	bibliografias XML catalogação
ARTIGO 17	Dublin Core bibliotecas digitais aplicações de computador XML metadados pré-publicações programas de computador programas de pós-graduação dissertações e teses bibliotecas
ARTIGO 18	grupos de discussão bibliotecas digitais treinamento manuais livros periódicos bibliografias pesquisa bibliotecas
ARTIGO 19	padrões de metadados acesso livre bibliotecas digitais sistemas de informação metadados normas e protocolos livros planejamento OAIS arquivos
ARTIGO 20	acesso remoto repositórios digitais tecnologias da informação e comunicação recursos de informação bibliotecas digitais terminologia ensino a distância comunicação acesso pesquisa
ARTIGO 21	indexação automática bibliotecas digitais sistemas de recuperação da informação entrada de dados bibliotecas centrais indexação recuperação da informação entradas

	resumos normas e protocolos
ARTIGO 22	comunidades científicas bibliotecas digitais disseminação da informação avaliação universidades papel educação bibliotecas dissertações e teses
ARTIGO 23	arquitetura de informação usos da informação bibliotecas digitais ciência da informação eficiência acesso pesquisa bibliotecas dissertações e teses
ARTIGO 24	bibliotecas virtuais bibliotecas digitais sociedade da informação tecnologias da informação e comunicação direito autoral Internet BD World Wide Web paradigmas livros
ARTIGO 25	competência em informação seleção de documentos acesso ao documento bibliotecas digitais estudantes ensino a distância acervos física bibliotecas educação
ARTIGO 26	acesso universal unidades de informação bibliotecas digitais acesso à informação cenários acervos história papel bibliotecas acesso

ARTIGO 27	perfil do usuário estudos de usuários bibliotecas digitais recuperação da informação questionários terminologia aplicativos empresas pesquisa navegação
ARTIGO 28	bibliotecas híbridas documentos eletrônicos bibliotecas virtuais organização do conhecimento transferência da informação fluxo da informação acesso à informação bibliotecas digitais acessibilidade bibliotecas
ARTIGO 29	bibliotecas virtuais bibliotecas digitais comunicação científica recursos de informação Internet questionários comunicação usuários pesquisadores papel
ARTIGO 30	acesso universal bibliotecas digitais cenários subsídios acesso bibliotecas
ARTIGO 31	bibliotecas digitais World Wide Web Internet conceitos bibliotecas
ARTIGO 32	bibliotecas universitárias bibliotecas virtuais bibliotecas digitais ciência da informação periódicos biblioteconomia comunicação autoria bibliotecas

ARTIGO 33	referências bibliográficas bibliotecas virtuais bibliotecas digitais tipos de documento desenvolvimento de coleções unidades de informação produtividade de autor artigos de periódico profissionais de informação planejamento
ARTIGO 34	documentos eletrônicos estudos de caso mineração de textos tecnologias da informação e comunicação bibliotecas digitais física algoritmos categorias World Wide Web catalogação
ARTIGO 35	bibliotecas nacionais bibliotecas digitais digitalização bibliotecas acervos
ARTIGO 36	preservação digital bibliotecas digitais gestão de bibliotecas serviços de informação competências profissionais profissionais de informação bibliotecas gestão papel gestores
ARTIGO 37	cooperação entre bibliotecas mecanismos de busca organização do conhecimento bibliotecas digitais acesso à informação custos cooperação bibliotecários Internet digitalização
ARTIGO 38	obras raras bibliotecas digitais acesso à informação acervos digitalização

	bibliotecas acesso
ARTIGO 39	preservação digital bibliotecas digitais bibliotecas universitárias equipamentos de computador digitalização programas de computador acervos acesso universidades bibliotecas
ARTIGO 40	bibliotecas universitárias programas livres bibliotecas digitais catálogos livros educação pesquisa bibliotecas programas de computador
ARTIGO 41	bibliotecas digitais redes de bibliotecas documentos eletrônicos registros bibliográficos publicações oficiais ciência da informação probabilidade e estatística registros publicações bibliotecas
ARTIGO 42	bibliotecas digitais repositórios institucionais acesso livre OAIS normas e protocolos cenários arquivos estudantes pesquisa bibliotecas
ARTIGO 43	bibliotecas digitais estudos de caso gestão do conhecimento gestão da informação indicadores de C&T indicadores empresas BD pesquisa

	gestão
ARTIGO 44	documentos eletrônicos bibliotecas digitais direito autoral ensino a distância estudantes manutenção acervos paradigmas gestão papel
ARTIGO 45	publicações eletrônicas comunicação científica artigos de periódico tecnologias da informação e comunicação recursos de informação acesso à informação bibliotecas digitais sistemas de informação necessidades de informação portais
ARTIGO 46	estudos de usuários buscas de informação redes de telecomunicações revisões de literatura bibliotecas digitais ciência da informação Internet tecnologias da informação e comunicação usabilidade psicologia
ARTIGO 47	preservação digital objetos digitais bibliotecas digitais normas e protocolos XML Metadata Encoding and Transmission Standard metadados acervos bibliotecas
ARTIGO 48	livros eletrônicos bibliotecas digitais multimídia ensino a distância lógica livros registros computadores Internet Educação

ARTIGO 49	classificação facetada modelos cognitivos redes de bibliotecas organização do conhecimento bibliotecas digitais recuperação da informação classificação mapas acervos termos
ARTIGO 50	bibliotecas digitais periódicos científicos tecnologias da informação e comunicação repositórios digitais informação científica e tecnológica comunicação científica acesso livre acesso à informação editores editoras
ARTIGO 51	arquitetura de informação bibliotecas digitais ciência da informação navegação buscas biblioteconomia bibliotecas
ARTIGO 52	bibliotecas universitárias ensino a distância serviços de biblioteca acesso livre bibliotecas digitais cooperação acesso universidades bibliotecas educação
ARTIGO 53	bibliotecas digitais informação científica e tecnológica ciência da informação cenários profissionais de informação paradigmas bibliotecários termos conceitos bibliotecas
ARTIGO 54	documentos eletrônicos recursos de informação bibliotecas digitais

	acessibilidade coleções conceitos livros bibliotecas
ARTIGO 55	pesquisa exploratória ensino de biblioteconomia bibliotecas digitais professores formação profissional bibliografias pesquisa estudantes levantamentos biblioteconomia
ARTIGO 56	direito autoral bibliotecas digitais tecnologias da informação e comunicação digitalização empresas livros direito bibliotecas conceitos pesquisa
ARTIGO 57	acesso remoto preservação digital obras raras tipos de documento preservação de documentos bibliotecas digitais acervos digitalização livros coleções
ARTIGO 58	bibliotecas digitais periódicos eletrônicos comportamento do usuário revisões de literatura informação científica e tecnológica disseminação da informação acesso à informação ciência da informação usabilidade portais
ARTIGO 59	revisões de literatura bibliotecários de referência unidades de informação serviços de referência bibliotecas digitais

	ciência da informação sistemas de informação profissionais de informação eficiência sistemas de recuperação da informação
ARTIGO 60	bibliotecas virtuais realidade virtual bibliotecas digitais paradigmas bibliotecas mudança
ARTIGO 61	representação da informação organização do conhecimento tecnologias da informação e comunicação bibliotecas digitais acesso livre recuperação da informação física metadados Código de Catalogação Anglo-Americano Indexação
ARTIGO 62	cultura organizacional periódicos científicos bibliotecas digitais projetos de pesquisa repositórios institucionais acesso livre portais bibliotecários cenários estudantes
ARTIGO 63	documentos eletrônicos bibliotecas digitais direito autoral direito à informação acesso à informação acervos digitalização Internet acesso pesquisa
ARTIGO 64	repositórios institucionais análise comparativa acesso livre bibliotecas digitais bases de dados probabilidade e estatística normalização gráficos acesso

	publicações
ARTIGO 65	ciências humanas acesso livre bibliotecas digitais manutenção livros programas de pós-graduação acesso pesquisa dissertações e teses usuários
ARTIGO 66	comunidades científicas periódicos científicos bibliotecas digitais acesso livre OAIS bibliometria metadados normas e protocolos livros periódicos
ARTIGO 67	pesquisa exploratória ensino a distância tecnologias da informação e comunicação bibliotecas digitais programas de pós-graduação dissertações e teses comunicação bibliotecas educação pesquisa
ARTIGO 68	estudos de caso livros eletrônicos bibliotecas digitais metadados funcionalidade livros Internet autoria dados usuários
ARTIGO 69	objetos digitais geoinformação acesso livre padrões de metadados redes de telecomunicações teorias na ciência da informação representação da informação tecnologias da informação e comunicação bibliotecas digitais

	ciência da informação
ARTIGO 70	bibliotecas digitais bibliotecas escolares estudos de caso organização do conhecimento sociedade da informação estados da arte artes gestão educação bibliotecas
ARTIGO 71	objetos digitais repositórios digitais documentos eletrônicos preservação digital bibliotecas digitais preservação de documentos metadados OAIS bibliotecas pesquisa
ARTIGO 72	repositórios digitais informação científica e tecnológica bibliotecas digitais acesso livre tecnologias da informação e comunicação recursos de informação similaridade livros estudantes história
ARTIGO 73	representação do conhecimento acesso à informação bibliotecas digitais ciência da informação sistemas de informação transdisciplinaridade hermenêutica títulos de documentos metadados epistemologia
ARTIGO 74	correio eletrônico serviços de referência bibliotecas digitais bibliotecários de referência tecnologias da informação e comunicação bibliotecários normas e protocolos bibliotecas

ARTIGO 75	bibliotecas digitais sistemas de recuperação da informação acesso à informação recuperação da informação taxonomias classificação dados buscas bibliotecas acesso
ARTIGO 76	serviços de biblioteca bibliotecas digitais propriedade intelectual sistemas de informação manutenção gestão bibliotecas pesquisa universidades dissertações e teses
ARTIGO 77	programas livres repositórios institucionais bibliotecas digitais programas de computador livros funcionalidade teste publicações bibliotecas universidades
ARTIGO 78	bibliotecas digitais ciências sociais aplicadas publicações seriadas artigos de periódico literatura cinzenta pequenas e médias empresas transferência da informação acesso à informação ciência da informação empresas
ARTIGO 79	bibliotecas digitais engenharia de produção documentos eletrônicos gestão do conhecimento bibliometria informetria professores engenharias arquivos estudantes

ARTIGO 80	usuários novatos buscas de informação bibliotecas digitais ciência da computação ciência da informação questionários sistemas de informação teste livros usabilidade
ARTIGO 81	linguagens documentárias bibliotecas virtuais bibliotecas digitais metadados acervos imagens medicina bibliotecas gestão educação
ARTIGO 82	economia da informação tecnologias da informação e comunicação serviços de biblioteca bibliotecas digitais transferência da informação economia mudança bibliotecas

APÊNDICE C – LISTA DE FREQUÊNCIA DE TÓPICOS INDEXADOS PELO MAUI NOS CORPORA

TÓPICO + FREQUÊNCIA – *CORPUS* A		TÓPICO + FREQUÊNCIA – *CORPUS* B	
ciência_da_informação	24	bibliotecas_digitais	82
pesquisa	17	biblioteca	39
educação	11	pesquisa	17
gestão	10	tecnologias_da_informação_e_comu	15
acesso	10	livro	15
profissionais_de_informação	9	acesso	15
usuário	8	ciência_da_informação	14
recuperação_da_informação	8	acesso_livre	13
papel	8	acesso_à_informação	12
gestão_da_informação	8	acervo	12
conceito	8	metadados	9
avaliação	8	educação	9
dado	7	dissertações_e_teses	9
comunicação	7	bibliotecas_universitárias	9
biblioteca	7	recuperação_da_informação	8
periódico	6	internet	8
internet	6	estudante	8
gestão_do_conhecimento	6	documentos_eletrônicos	8
empresa	6	digitalização	8
autoria	6	bibliotecas_virtuais	8
universidade	5	sistemas_de_informação	7
tecnologias_da_informação_e_com		preservação_digital	7
unicação	5	gestão	7
sistemas_de_informação	5	bibliotecário	7
inteligência_competitiva	5	universidade	6
world_wide_web	4	papel	6
publicação	4	organização_do_conhecimento	6
inovação	4	normas_e_protocolos	6
informação_científica_e_tecnológica	4	estudos_de_caso	6
hipertexto	4	ensino_a_distância	6
gestor	4	direito_autoral	6
formação_profissional	4	usuário	5
economia	4	repositórios_institucionais	5
confiabilidade	4	recursos_de_informação	5
comunicação_científica	4	programas_de_computador	5
biblioteconomia	4	objetos_digitais	5
bibliometria	4	informação_científica_e_tecnológica	5
bases_de_dados	4	conceito	5
áreas_do_conhecimento	3	comunicação	5
transferência_da_informação	3	cenário	5
termo	3	usabilidade	4

Sociologia	3	unidades_de_informação	4
sociedade_da_informação	3	sistemas_de_recuperação_da_informação	4
Planejamento	3	revisões_de_literatura	4
Pesquisador	3	repositórios_digitais	4
normas_e_protocolos	3	questionário	4
necessidades_de_informação	3	publicação	4
Mudança	3	profissionais_de_informação	4
Levantamento	3	professor	4
dissertações_e_teses	3	paradigma	4
Direito	3	oais	4
Busca	3	gestão_da_informação	4
Arquivo	3	empresa	4
usos_da_informação	2	disseminação_da_informação	4
tomada_de_decisões	2	coleção	4
serviços_de_empréstimo	2	busca	4
revisões_de_literatura	2	bibliografia	4
relevância	2	arquivo	4
Registro	2	arquitetura_de_informação	4
referências_bibliográficas	2	xml	3
recursos_de_informação	2	world_wide_web	3
propriedade_intelectual	2	transferência_da_informação	3
programas_de_computador	2	sociedade_da_informação	3
produtividade_de_autor	2	serviços_de_biblioteca	3
processamento_de_informações	2	redes_de_telecomunicações	3
políticas_de_informação	2	redes_de_bibliotecas	3
periódicos_eletrônicos	2	programas_de_pós	3
paradigma	2	portar	3
monitoramento	2	periódicos_científicos	3
mercado_de_trabalho	2	periódico	3
informação_para_negócios	2	manutenção	3
informação_governamental	2	livros_eletrônicos	3
identificadores_de_objetos_digitais	2	graduação	3
História	2	físico	3
governo_eletrônico	2	funcionalidade	3
fluxo_da_informação	2	eficiência	3
Filosofia	2	desenvolvimento_de_coleções	3
Estudante	2	cooperação	3
estratégias_de_busca	2	comunicação_científica	3
direito_à_privacidade	2	biblioteconomia	3
direito_autoral	2	bibliometria	3
conhecimento_nas_organizações	2	artigos_de_periódico	3
conceitos_de_informação	2	tipos_de_documento	2
computador	2	teste	2
complexidade	2	termo	2
competência_em_informação	2	terminologia	2
competitividade	2		

classificação	2	subsídio	2
cenário	2	serviços_de_referência	2
categoria	2	representação_da_informação	2
buscas_de_informação	2	registro	2
bibliotecário	2	propriedade_intelectual	2
bibliotecas_virtuais	2	projetos_de_pesquisa	2
bibliotecas_digitais	2	programas_livres	2
artigos_de_periódico	2	probabilidade_e_estatística	2
agências_de_fomento	2	preservação_de_documentos	2
administração	2	planejamento	2
índice_h	1	pesquisa_exploratória	2
índice	1	periódicos_eletrônicos	2
xml	1	padrões_de_metadados	2
validade	1	obras_raras	2
usabilidade	1	necessidades_de_informação	2
unidades_de_informação	1	navegação	2
títulos_de_documentos	1	mudança	2
tipos_de_documento	1	mapa	2
tesauros	1	levantamento	2
teorias_na_ciência_da_informação	1	inovação	2
teoria_do_caos	1	indicadores_de_c	2
sumário	1	indicador	2
subsídio	1	indexação	2
sistemas_de_recuperação_da_infor mação	1	imagem	2
similaridade	1	história	2
semântica	1	gestão_do_conhecimento	2
semiótica	1	estudos_de_usuários	2
relações_entre_conceitos	1	entrevista	2
regimes_de_informação	1	engenharia	2
publicações_eletrônicas	1	economia_da_informação	2
projetos_de_pesquisa	1	direito_à_informação	2
programas_de_pós	1	dado	2
professor	1	cooperação_entre_bibliotecas	2
processos_de_gestão	1	comunidades_científicas	2
probabilidade_e_estatística	1	classificação	2
pesquisa_e_desenvolvimento	1	ciências_sociais_aplicadas	2
pertinência	1	categoria	2
periódicos_científicos	1	catalogação	2
orçamento	1	buscas_de_informação	2
organização_do_conhecimento	1	bibliotecários_de_referência	2
nutrição	1	bibliotecas_centrais	2
notícia	1	bd	2
nomes_próprios	1	bases_de_dados	2
navegação	1	avaliação	2
métodos_de_pesquisa	1	autoria	2
		acesso_universal	2

multidisciplinaridade	1	acesso_remoto	2
momografia	1	acessibilidade	2
medicina	1	usuários_novatos	1
mecanismos_de_busca	1	usos_da_informação	1
livro	1	títulos_de_documentos	1
literatura_cinzenta	1	treinamento	1
linguagens_documentárias	1	transdisciplinaridade	1
linguagens_de_marcação	1	tomada_de_decisões	1
lei_de_lotka	1	teorias_na_ciência_da_informação	1
lei_de_bradford	1	teoria_do_conceito	1
intranetes	1	taxonomia	1
interdisciplinaridade	1	superposição	1
inteligência_artificial	1	similaridade	1
institutos_de_pesquisa	1	serviços_de_informação	1
informetria	1	seleção_de_documentos	1
informação_jurídica	1	resumo	1
informação_financeira	1	representação_do_conhecimento	1
indústria_da_informação	1	registros_bibliográficos	1
indicadores_de_c	1	referências_bibliográficas	1
indicador	1	realidade_virtual	1
indexação	1	publicações_seriadas	1
inclusão_digital	1	publicações_oficiais	1
imagem	1	publicações_eletrônicas	1
html	1	psicologia	1
história_da_ciência_da_informação	1	produtividade_de_autor	1
guia	1	predeterminado	1
gráfico	1	políticas_públicas	1
graduação	1	pesquisador	1
físico	1	perfil_do_usuário	1
frente_de_pesquisa	1	pequenas_e_médias_empresas	1
formatos_marc	1	normalização	1
estudos_experimentais	1	multimídia	1
estudos_de_usuários	1	modelos_cognitivos	1
estudos_de_caso	1	mineração_de_textos	1
estruturalismo	1	metadata_encoding_and_transmission_standard	1
estados_da_arte	1	medicina	1
ensino_a_distância	1	mecanismos_de_busca	1
educação_superior	1	manual	1
editora	1	lógica	1
editor	1	literatura_cinzenta	1
economia_da_informação	1	linguagens_documentárias	1
ecologia	1	informetria	1
divulgação_científica	1	informação_governamental	1
disseminação_seletiva_da_informação	1	indexação_automática	1
disseminação_da_informação	1	hermenêutica	1

discos_compactos	1	gráfico	1
direito_à_informação	1	grupos_de_discussão	1
desenvolvimento_de_coleções	1	governo_eletrônico	1
Descritor	1	gestão_de_bibliotecas	1
Descarte	1	gestores_de_bibliotecas	1
Custo	1	gestor	1
cooperação	1	geoinformação	1
controle_bibliográfico	1	formação_profissional	1
Consultor	1	fluxo_da_informação	1
ciências_da_saúde	1	estágio	1
ciência_da_computação	1	estados_da_arte	1
ciência_cognitiva	1	equipamentos_de_computador	1
citações_bibliográficas	1	epistemologia	1
cientometria	1	entrada_de_dados	1
cientistas_da_informação	1	entrada	1
Cientista	1	ensino_de_biblioteconomia	1
Carta	1	engenharia_de_produção	1
bibliotecas_híbridas	1	engenharia_biomédica	1
autoria_individual	1	editora	1
Arte	1	editor	1
Arranjo	1	economia	1
arquivos_privados	1	dublin_core	1
arquivologia	1	documentos_primários	1
aplicações_de_computador	1	direito	1
análise_de_citação	1	dados_científicos	1
ambiente_organizacional	1	código_de_catalogação_anglo	1
agentes_inteligentes	1	custo	1
acesso_à_informação	1	cultura_organizacional	1
acesso_livre	1	correio_eletrônico	1
acessibilidade	1	controle_bibliográfico	1
acervo	1	computador	1
		comportamento_do_usuário	1
		competências_profissionais	1
		competência_em_informação	1
		classificação_facetada	1
		classificação_automática	1
		ciências_humanas	1
		ciência_da_computação	1
		catálogo	1
		brecha_digital	1
		bibliotecas_nacionais	1
		bibliotecas_híbridas	1
		bibliotecas_escolares	1
		arte	1
		aplicações_de_computador	1
		aplicativo	1

	análise_qualitativa	1
	análise_comparativa	1
	americano	1
	algoritmo	1
	acesso_ao_documento	1

APÊNDICE D – COMPARAÇÃO ENTRE AS PALAVRAS-CHAVE E OS TÓPICOS INDEXADOS – *CORPUS* A

Palavras-Chave do *Corpus* A	Descritores Atribuídos pelo Indexador Manual	Tópicos atribuídos pelo MAUI do *Corpus* A
ARTIGO 01 Transferência de informação; Gestão do conhecimento; Valor de unidades de conhecimento	transferência da informação gestão do conhecimento gestão da informação conhecimento nas organizações acesso à informação competitividade conhecimento tácito gestão de documentos repositórios digitais processos de gestão disseminação da informação	conhecimento nas organizações transferência da informação gestão do conhecimento descarte avaliação acesso gestão
ARTIGO 02 Popularização da Ciência; Comunicação Científica	divulgação científica comunicação científica ciência da informação educação notícias acesso à informação áreas do conhecimento gestão da informação sociedade da informação informação científica e tecnológica	comunicação científica divulgação científica ciência da informação notícias educação comunicação
ARTIGO 03 Informação; Valor Informacional; Direito à Informação; Memória Social; Estoque Informacional	direito à informação recuperação da informação acesso à informação ciência da informação sociedade da informação atributos da informação ética na informação biblioteconomia história liberdade de pensamento	direito à informação recuperação da informação avaliação direito confiabilidade confiabilidade
ARTIGO 04 Arquivos-abertos; Sistema de Publicação; Budapest Open Access Initiative; Acesso Livre; Auto-arquivamento	acesso livre autoarquivamento de documentos comunicação científica publicações científicas publicações eletrônicas periódicos científicos direito autoral literatura científica repositórios digitais tecnologias da informação e comunicação	acesso livre direito autoral arquivos livros filosofia conceitos publicações arranjo direito acesso
ARTIGO 05 Conhecimento Científico; Conhecimento Privado; Conhecimento Escolar; Democratização da Ciência; Comunicação Científica	divulgação científica comunicação científica informação cientifica e tecnológica acesso à informação sociedade da informação educação estudantes áreas do conhecimento comunidades científicas dados científicos	comunicação científica informação científica e tecnológica sumários educação estudantes cientistas acesso comunicação pesquisa
ARTIGO 06	interdisciplinaridade ciência da computação	ciência da computação ciência da informação

Tesauro Eletrônico; Mundo do Trabalho; Recuperação da Informação; Interface de Consulta; Sistema de Informação; Interdisciplinaridade; Interação Humano-Computador	ciência da informação manutenção de tesauros tesauros elaboração de linguagens documentárias recuperação da informação interação homem-computador sistemas de informação mercado de trabalho	interdisciplinaridade recuperação da informação sistemas de informação navegação computadores gestão pesquisa tesauros
ARTIGO 07 Inteligência Competitiva; Gestão do Conhecimento; Gestão da Informação; Fluxos Informacionais; Transferência da Informação	inteligência competitiva gestão do conhecimento gestão da informação fluxo da informação transferência da informação competitividade ciência da informação economia da informação informação para negócios conhecimento nas organizações recursos de informação	transferência da informação recursos de informação gestão do conhecimento inteligência competitiva gestão da informação competitividade papel dados gestão
ARTIGO 08 Informação; Massa Documental; Conceito de Informação; Tecnologia; Registro do Conhecimento	conceitos de informação ciência da informação áreas do conhecimento tecnologias da informação e comunicação indústria da informação conteúdos da informação disseminação da informação explosão da informação gestão de documentos gestão do conhecimento modelagem do conhecimento registros bibliográficos sociedade da informação tipos de documento	conceitos de informação estados da arte ciência da informação áreas do conhecimento homografia artes registros pesquisa termos conceitos
ARTIGO 09 Universidade; Gestão do fluxo de Informação na Universidade; Inteligência Competitiva; Barreiras na Comunicação da Informação; Pecados Informacionais	fluxo da informação universidades inteligência competitiva gestão da informação tomada de decisões ciência da informação tecnologias da informação e comunicação comunicação nas organizações cultura organizacional informação estratégica obsolescência da literatura comunicação informal	conceitos de informação tomada de decisões disseminação seletiva da informação fluxo da informação ciência da informação inteligência competitiva gestão gestores universidades conceitos
ARTIGO 10 Informação Estatística; Nova Economia; Mensuração Estatística; Desajuste Conceitual; Metodologia Estatística	economia métodos matemáticos e estatísticos probabilidade e estatística levantamentos dados numéricos informação governamental sociedade da informação tecnologias da informação e comunicação atributos da informação	probabilidade e estatística pertinência levantamentos economia

	inteligência competitiva sociedade do conhecimento sistemas de informação	
ARTIGO 11 Ciência da Informação; Armazenamento e recuperação; Curso em informação; Unidade e especificidade da informação	ciência da informação ensino de ciência da informação recuperação da informação armazenamento de dados educação pesquisa gestão da informação interdisciplinaridade propriedade intelectual unidades de informação	ciência da informação propriedade intelectual recuperação da informação educação classificação pesquisa dados
ARTIGO 12 Ciência da Informação; Biblioteconomia; Sistema de Informação; Arquivologia; Ensino; Pesquisa	ciência da informação biblioteconomia arquivologia sistemas de informação ensino de ciência da informação profissionais de informação acesso à informação educação pesquisa teoria da informação	acesso à informação arquivologia sistemas de informação áreas do conhecimento ciência da informação nomes próprios acesso biblioteconomia educação pesquisa
ARTIGO 13 Profissional da informação; Informação organizacional; Formação e profissional da informação	formação profissional profissionais de informação ciência da informação competências profissionais conhecimento nas organizações disseminação da informação economia gestão do conhecimento sociedade da informação sistemas de informação bibliotecas biblioteconomia profissão e mercado de trabalho	profissionais de informação sistemas de informação formação profissional bibliotecas economia usuários papel acervos
ARTIGO 14 Profissionais da informação; Funções sociais; Perfis de profissionais da informação; Inclusão digital; Gestão da informação; Gestão do conhecimento	profissionais de informação formação profissional competências profissionais gestão da informação gestão do conhecimento sociedade da informação mercado de trabalho fluxo da informação inclusão digital competência em informação engenharia do conhecimento profissão e mercado de trabalho	inclusão digital competência em informação gestão do conhecimento profissionais de informação gestão da informação necessidades de informação gestão acesso papel gestores
ARTIGO 15 Formação profissional; Ensino e pesquisa	profissionais de informação formação profissional competências profissionais ensino de ciência da informação ensino e pesquisa em ciência da informação e áreas afins educação superior pesquisa ciência da informação paradigmas instituições de ensino e pesquisa sociedade do conhecimento	formação profissional profissionais de informação mudança universidades pesquisa educação

ARTIGO 16 Ciência da Informação; Formação profissional; Educação Superior no Brasil; Sociedade da Informação; educação; Ciência da Informação e Biblioteconomia; Ciência da Informação; curso de graduação	ciência da informação educação superior formação profissional biblioteconomia sociedade da informação ensino de ciência da informação ensino de biblioteconomia ensino e pesquisa em ciência da informação e áreas afins profissionais de informação competências profissionais cientistas da informação	educação superior sociedade da informação formação profissional ciência da informação profissionais de informação mudança biblioteconomia educação
ARTIGO 17 Internet e Produção Científica; Novas Tecnologias de Informação e de Comunicação; Produção Científica e Novas Tecnologias	comunicação científica publicações eletrônicas Internet tecnologias da informação e comunicação inovação sociedade da informação produtividade científica armazenamento de dados controle bibliográfico propriedade intelectual segurança da informação	controle bibliográfico tecnologias da informação e comunicação publicações eletrônicas propriedade intelectual direito à privacidade comunicação publicações relevância autoria complexidade
ARTIGO 18 Estudos Sociais da Ciência; Sociologia da Ciência; Ciência e Sociedade; Tecnologia e Sociedade	sociedade da informação tecnologias da informação e comunicação pesquisa divulgação científica categorias construtivismo taxonomias ciência da informação economia da informação escolas e correntes filosóficas produtividade cientifica relações entre conceitos sociedades científicas sociologia do conhecimento sociologia	sociologia categorias
ARTIGO 19 Inteligência Empresarial; Monitoração Ambiental; Fontes de Informação; Gestão do Conhecimento; Gestão da Informação	inteligência competitiva monitoramento ambiental gestão da informação gestão do conhecimento ambiente organizacional conhecimento nas organizações informação para negócios recursos de informação bases de dados bibliotecas especializadas competitividade relevância tecnologias da informação e comunicação	ambiente organizacional gestão do conhecimento gestão da informação inteligência competitiva avaliação bibliotecas gestão pesquisa confiabilidade confiabilidade
ARTIGO 20 Ciência da Informação no Brasil; Avaliação 2001 CAPES; Pós-graduação em Ciência da Informação;	ciência da informação programas de pós-graduação em ciência da informação pesquisa em ciência da informação projetos de pesquisa avaliação de publicações científicas	ciência da informação gráficos publicações pesquisa dissertações e teses avaliação

Pesquisa em Ciência da Informação no Brasil	dissertações e teses ensino de ciência da informação profissionais de informação programas de pós-graduação agências de fomento publicações	
ARTIGO 21 Leitura-teoria; Cognóscio; Conhecimento-introjeção; Leitura e Sociedade; Informação e Sociedade	sociedade da informação promoção do livro e da leitura ciência cognitiva inteligência artificial linguística sociologia multidisciplinaridade áreas do conhecimento usuários	inteligência artificial áreas do conhecimento multidisciplinaridade sociologia autoria usuários dissertações e teses
ARTIGO 22 Paradigma; Holografia, Ciência da Informação; Tecnologia da Informação; Hipertexto; Complexidade; Interatividade; Virtual; Totalidade	paradigmas ciência da informação tecnologias da informação e comunicação hipertextos complexidade organização da informação processamento de informações	tecnologias da informação e comunicação processamento de informações ciência da informação hipertextos hipertexto formatos MARC serviços de empréstimo paradigmas pesquisa biblioteconomia
ARTIGO 23 Informação; Memória Social; Espaço Prisional	disseminação da informação ciência da informação documentação história sociedade da informação interdisciplinaridade relações entre conceitos teoria da informação tipos de documento conceitos de informação bibliotecas de prisões	relações entre conceitos disseminação da informação ciência da informação história
ARTIGO 24 Contrato Social; Ciência; Pesquisa; Pesquisadores; Autonomia; Ecologia dos Conhecimentos	pesquisa pesquisadores políticas públicas comunicação científica informação cientifica e tecnológica institutos de pesquisa universidades paradigmas transdisciplinaridade áreas do conhecimento	ecologia validade buscas pesquisadores pesquisa
ARTIGO 25 Fomento à pesquisa - Ciência da Informação; CNPq - fomento à pesquisa em Ciência da Informação	agências de fomento ciência da informação bolsas de pesquisa projetos de pesquisa gestão de recursos humanos pesquisa pesquisadores pesquisa e desenvolvimento programas de pós-graduação em ciência da informação levantamentos dados numéricos informação financeira	ciência da informação agências de fomento programas de pós-graduação orçamento levantamentos pesquisadores pesquisa dados

	orçamento	
ARTIGO 26 Compressão Semântica; Monitoramento da Informação; Estoques de Informação; Palavras-chave	monitoramento palavras-chaves semântica compressão de dados informação estratégica recuperação da informação processamento de textos processamento da linguagem natural relevância	ciência da informação programas de computador subsídios semântica administração monitoramento gestão
ARTIGO 27 Indicadores Científicos; Política Científica e Tecnológica; Gestão de Ciência e Tecnologia	indicadores de C&T bolsas de pesquisa agências de fomento pesquisadores bases de dados bases de dados referenciais análise de dados produtividade cientifica projetos de pesquisa	bases de dados monitoramento agências de fomento indicadores de C&T indicadores gestão relevância pesquisadores dados
ARTIGO 28 Governo Eletrônico; Arquitetura de Sistemas de Informação; Integração de Informações; Gestão de C&T; Bibliotecas Digitais, Plataforma Lattes; Rede SCienTI	sistemas de informação governo eletrônico serviços de informação gestão da informação unidades de informação bibliotecas digitais inovação tecnologias da informação e comunicação interoperabilidade	bibliotecas digitais informação governamental gestão da informação necessidades de informação governo eletrônico sistemas de informação inovação bibliotecas papel dissertações e teses
ARTIGO 29 Governo Eletrônico; Políticas de Informação; Informação Governamental	governo eletrônico políticas de informação informação governamental acesso à informação gestão da informação gestão de conteúdos na web serviços de informação	políticas de informação governo eletrônico gestão da informação informação governamental gestão identificadores de objetos digitais acesso administração
ARTIGO 30 Periódicos eletrônicos; Avaliação de acesso; Arquivo de log de acesso	análise de dados periódicos eletrônicos periódicos científicos artigos de periódico world wide web registros de uso coleta de dados estudos de usuários perfil do usuário usuários disseminação da informação processos de gestão publicações de acesso livre recuperação da informação	periódicos eletrônicos World Wide Web avaliação periódicos arquivos acesso
ARTIGO 31 Política de informação; Sociedade da informação; Internet; Institucionalização da informação; Estado	políticas de informação sociedade da informação Internet infraestrutura de informação economia economia da informação governo eletrônico regimes de informação relações entre conceitos	regimes de informação sociedade da informação políticas de informação Internet serviços de empréstimo cenários semiótica conceitos economia

		termos
ARTIGO 32 Bases de dados; Estratégia de busca; Linguagem controlada; Linguagem natural. Recuperação da informação; Artigo de revisão	bases de dados estratégias de busca linguagens documentárias recuperação da informação revisões de literatura controle de vocabulário processamento da linguagem natural termos de indexação	linguagens documentárias buscas de informação revisões de literatura estratégias de busca bases de dados recuperação da informação planejamento discos compactos termos complexidade
ARTIGO 33 Biblioteca digital; Biblioteca virtual; Produção científica; Produção bibliográfica; Periódicos	bibliotecas digitais bibliotecas virtuais artigos de periódico citações bibliográficas informação cientifica e tecnológica produtividade científica produtividade de periódicos produtividade de autor revisões de literatura tecnologias da informação e comunicação profissionais de informação bibliotecas Internet	referências bibliográficas bibliotecas virtuais bibliotecas digitais tipos de documento desenvolvimento de coleções unidades de informação produtividade de autor artigos de periódico profissionais de informação planejamento
ARTIGO 34 Informação tecnológica; Transferência de informação; Transferência tecnológica	informação científica e tecnológica transferência da informação inovação acesso à informação ciência e tecnologia de alimentos nutrição comunicação científica fluxo da informação interação universidade-empresa pesquisa e desenvolvimento tecnologias da informação e comunicação institutos de pesquisa universidades	ciências da saúde informação científica e tecnológica transferência da informação nutrição mudança acesso comunicação universidades
ARTIGO 35 XML; HTML; Linguagens de marcação; Internet; Intranet	linguagens de marcação Internet intranetes XML HTML SGML ciência da informação	linguagens de marcação ciência da informação intranetes HTML XML editores normas e protocolos editoras publicações Internet
ARTIGO 36 Bibliometria; Lei de Lotka; Produtividade de autores; Brasil	bibliometria lei de Lotka produtividade de autor disseminação da informação literatura científica métodos matemáticos e estatísticos bibliografias	arquivos privados produtividade de autor lei de Lotka cartas periódicos medicina arquivos bibliometria autoria dados

ARTIGO 37 Informação para negócios; Bases de dados	informação para negócios bases de dados tecnologias da informação e comunicação tomada de decisões acesso à informação empresas recursos de informação sociedade da informação informação financeira informação jurídica bases de dados factuais indústria da informação	informação jurídica informação financeira indústria da informação informação para negócios tomada de decisões bases de dados guias classificação cenários empresas
ARTIGO 38 Biblioteca híbrida; Tipos de usuários; Bens e serviços	bibliotecas híbridas ensino a distância usuários serviços de informação acesso à informação necessidades de informação tecnologias da informação e comunicação Internet bibliotecas universitárias bibliotecas digitais recursos de informação serviços de biblioteca suportes de informação usos da informação	bibliotecas híbridas tecnologias da informação e comunicação ensino a distância usuários Internet papel comunicação universidades bibliotecas educação
ARTIGO 39 Estratégia de busca; Recuperação da informação; Técnicas de estratégia de busca; Bases de dados; Artigo de revisão	estratégias de busca recuperação da informação bases de dados revisões de literatura sistemas de recuperação da informação mecanismos de busca interdisciplinaridade	sistemas de recuperação da informação buscas de informação revisões de literatura estratégias de busca bases de dados recuperação da informação planejamento acesso conceitos usuários
ARTIGO 40 Ciência da informação; Gestão da informação	gestão da informação ciência da informação profissionais de informação formação profissional recursos de informação mercado de trabalho competências profissionais tecnologias da informação e comunicação tomada de decisões explosão da informação gestão do conhecimento necessidades de informação	direito autoral mercado de trabalho recursos de informação ciência da informação profissionais de informação gestão da informação direito à privacidade direito gestores dados
ARTIGO 41 Produção científica; Literatura branca; Literatura cinzenta; Ciência da informação	ciência da informação programas de pós-graduação programas de pós-graduação em ciência da informação literatura científica literatura cinzenta artigos de periódico bibliometria	autoria individual literatura cinzenta artigos de periódico institutos de pesquisa ciência da informação índices índice h periódicos

	produtividade científica comunidades científicas história da ciência da informação publicações	autoria dados
ARTIGO 42 Gestão do conhecimento; Capital intelectual; Informação para negócios; Sistemas de informação para negócios; Agentes do conhecimento	informação para negócios gestão do conhecimento gestão da informação sistemas de informação consultores de informação inteligência competitiva bibliotecas bibliotecas virtuais profissionais de informação serviços de biblioteca unidades de informação empresas	bibliotecas virtuais informação para negócios gestão do conhecimento sistemas de informação bibliotecas gestão da informação inteligência competitiva profissionais de informação empresas usuários
ARTIGO 43 Necessidade de informação tecnológica; Informação tecnológica; Setor industrial; Inovação	informação científica e tecnológica inovação necessidades de informação tomada de decisões conhecimento nas organizações empresas análise de informação na indústria cinco forças de porter	informação científica e tecnológica necessidades de informação empresas inovação pesquisa
ARTIGO 44 Gestão do conhecimento; Informação e competitividade; Processos organizacionais	competitividade conhecimento nas organizações gestão do conhecimento informação estratégica processos de gestão ambiente organizacional inovação tomada de decisões tecnologias da informação e comunicação métodos de análise na inteligência competitiva ensino e pesquisa em ciência da informação e áreas afins	referências bibliográficas tecnologias da informação e comunicação gestão do conhecimento conhecimento nas organizações empresas competitividade papel gestão
ARTIGO 45 Bibliometria; Cienciometria; Informetria; Webometria; Métodos quantitativos de avaliação	bibliometria informetria cientometria webmetria análise quantitativa métodos matemáticos e estatísticos bases de dados ciência da informação	fluxo da informação ciência da informação Internet registros bibliometria similaridade informetria avaliação comunicação autoria
ARTIGO 46 bibliometria; Lei de Bradford; Pesquisa operacional; Caos; Ciência da informação; Inferência bayesiana	bibliometria lei de Bradford ciência da informação periódicos teoria do caos modelos matemáticos métodos matemáticos e estatísticos	métodos de pesquisa lei de Bradford teoria do caos ciência da informação bibliometria física periódicos conceitos pesquisa identificadores de objetos digitais
ARTIGO 47	periódicos eletrônicos periódicos científicos	periódicos científicos periódicos eletrônicos

Periódicos eletrônicos; Usabilidade; Novas tecnologias	usuários World Wide Web ciência da informação usabilidade estudos de usuários buscas de informação buscas em linha indexação automática usos da informação economia da informação	ciência da informação usabilidade usuários periódicos World Wide Web
ARTIGO 48 Uso da informação; Economia da Informação; Modelo genérico	usos da informação economia da informação inovação inteligência competitiva custos empresas gestão da informação informação para negócios tecnologias da informação e comunicação	usos da informação economia da informação custos inovação empresas buscas levantamentos economia
ARTIGO 49 Monitoramento da informação; Biblioteconomia; Ciência da informação	monitoramento bibliometria biblioteconomia ciência da informação periódicos descritores pesquisadores inteligência competitiva gestão do conhecimento bases de dados periódicos científicos	pesquisa e desenvolvimento ciência da informação descritores periódicos títulos de documentos programas de computador World Wide Web pesquisa biblioteconomia autoria
ARTIGO 50 Educação dos bibliotecários; Profissional da informação	ensino de biblioteconomia profissionais de informação bibliotecários biblioteconomia formação profissional educação sociedade da informação gestão da informação gestão do conhecimento	cooperação formação profissional profissionais de informação bibliotecários conceitos papel educação
ARTIGO 51 Acessibilidade; Espaço digital; Bibliotecas; Pessoas portadoras de deficiência; Ajudas técnicas	acessibilidade bibliotecas acesso à informação usuários perfil do usuário comunicação mediada por computador sistemas de informação tecnologias da informação e comunicação inclusão digital Internet digitalização	acessibilidade categorias usuários bibliotecas
ARTIGO 52 Estudos de usuários; Educação ambiental; Internet; Hipertexto; Pesquisa participante	estudos de usuários ciência da informação usuários Internet projetos de pesquisa educação hipertextos disseminação da informação	projetos de pesquisa estudos de usuários ciência da informação normas e protocolos hipertextos hipertexto Internet papel

	fluxo da informação	educação pesquisa
ARTIGO 53 Análise de logs; Máquinas de Busca; Recuperação da Informação; Comportamento de usuários; Estratégia de busca	comportamento do usuário estratégias de busca mecanismos de busca recuperação da informação linguagens documentárias buscas de informação serviços de referência usuários	mecanismos de busca recuperação da informação estudos experimentais professores estudantes World Wide Web usuários buscas universidades
ARTIGO 54 Information literacy; Competência em informação; Alfabetização informacional; Biblioteca aprendente; Bibliotecário educador; Sociedade de aprendizagem; Habilidades informacionais	competência em informação bibliotecários bibliotecas sociedade da informação educação buscas de informação recursos de informação gestão do conhecimento usos da informação profissionais de informação	sociedade da informação competência em informação paradigmas filosofia bibliotecários educação bibliotecas avaliação acesso conceitos
ARTIGO 55 Profissional da informação; Profissional da informação – habilidades; Perfil e atuação profissional; Mercado de trabalho	competências profissionais profissionais de informação mercado de trabalho formação profissional ensino de ciência da informação inteligência competitiva ciência da informação sistemas de informação conhecimento nas organizações	mercado de trabalho profissionais de informação ciência da informação imagens empresas consultores gestores
ARTIGO 56 Teoria da ciência da informação; Sociologia da informação; História da ciência da informação; Comunicação científica; Responsabilidade social	teorias na ciência da informação história da ciência da informação comunicação científica ciência da informação cientistas da informação biblioteconomia informação cientifica e tecnológica interdisciplinaridade sociologia epistemologia da ciência da informação	comunicação científica teorias na ciência da informação cientistas da informação organização do conhecimento história da ciência da informação informação científica e tecnológica ciência da informação normas e protocolos sociologia história
ARTIGO 57 Recuperação da informação; Inteligência científica; Integração dos conhecimentos; Estado; Ciência; Sociedade; Informação	ciência da informação recuperação da informação sociedade da informação paradigmas regimes de informação pesquisa áreas do conhecimento ensino e pesquisa em ciência da informação e áreas afins epistemologia da ciência da informação	recuperação da informação ciência da informação pesquisa avaliação
ARTIGO 58 Ciência da informação; Ciência cognitiva; Processamento da informação; Categorização; Indexação; Recuperação da informação; Interação	ciência da informação ciência cognitiva processamento de informações indexação recuperação da informação interação homem-computador categorização automática de textos	ciência cognitiva processamento de informações recuperação da informação ciência da informação indexação computadores

homem-computador		
ARTIGO 59 Comunicação científica; Bibliometria; Comunicação e Educação; Estudo de citações; Cientometria	comunicação científica bibliometria cientometria análise de citação educação comunicação frente de pesquisa padrões de comunicação científica epistemologia arqueológica produtividade cientifica	comunicação científica frente de pesquisa citações bibliográficas análise de citação estruturalismo cientometria bibliometria comunicação educação pesquisa
ARTIGO 60 Inteligência competitiva; Internet; Monitoramento de fontes de informação; Agentes inteligentes	agentes inteligentes inteligência competitiva internet monitoramento gestão da informação acesso à informação aplicações de computador programas de computador estudos de caso inovação usos da informação disseminação da informação competitividade	agentes inteligentes estudos de caso processos de gestão tecnologias da informação e comunicação usos da informação aplicações de computador gestão da informação inteligência competitiva inovação Internet

APÊNDICE E – COMPARAÇÃO ENTRE AS PALAVRAS-CHAVE E OS TÓPICOS INDEXADOS – *CORPUS* B

Palavras-Chave do *Corpus* B	Descritores Atribuídos pelo Indexador Manual do *Corpus* B	Tópicos atribuídos pelo MAUI do *Corpus* B
ARTIGO 01 biblioteca digital; teses on-line	bibliotecas digitais dissertações e teses universidades	bibliotecas digitais funcionalidade bibliotecas universidades dissertações e teses
ARTIGO 02 tecnologias da informação e comunicação; racismo; conteúdo freireano; biblioteca digital; educação	bibliotecas digitais tecnologias da informação e comunicação educação	políticas públicas projetos de pesquisa tecnologias da informação e comunicação bibliotecas digitais imagens educação pesquisa comunicação acesso bibliotecas
ARTIGO 03 biblioteca digital; biblioteca universitária; ciência eletrônica; desenvolvimento de coleções; ensino superior; gestão da informação; internet; livro eletrônico; periódico eletrônico; referência digital; repositório eletrônico; repositório de dados científicos; universidade	bibliotecas digitais bibliotecas universitárias ensino superior repositórios digitais livros eletrônicos periódicos eletrônicos dados científicos	cooperação entre bibliotecas dados científicos bibliotecas universitárias bibliotecas digitais periódicos eletrônicos desenvolvimento de coleções gestores de bibliotecas unidades de informação livros eletrônicos gestão da informação
ARTIGO 04 biblioteca digital; preservação digital; universidade de Brasília; biblioteca universitária	bibliotecas digitais bibliotecas universitárias preservação digital	bibliotecas universitárias bibliotecas centrais preservação digital bibliotecas digitais análise qualitativa revisões de literatura entrevistas gestão buscas conceitos
ARTIGO 05 biblioteca digital; ciência da informação; responsabilidade social; sistema de recuperação da informação	bibliotecas digitais sistemas de recuperação da informação objetos digitais disseminação da informação recuperação da informação ciência da informação	objetos digitais sistemas de recuperação da informação bibliotecas digitais disseminação da informação recuperação da informação ciência da informação mapas usuários bibliotecas
ARTIGO 06	bibliotecas digitais	documentos primários

Bibliotecas digitais; Metadados; Ontologias; Teoria do conceito; Bibliotecas	teoria do conceito metadados ontologias recuperação da informação classificação automática	objetos digitais classificação automática teoria do conceito bibliotecas digitais ciência da informação recuperação da informação bibliotecários bibliometria professores
ARTIGO 07 biblioteca digital; brecha digital; países em desenvolvimento; indicadores de impacto; cooperação interbibliotecária	bibliotecas digitais brecha digital bibliotecas universitárias cooperação entre bibliotecas indicadores	brecha digital bibliotecas universitárias bibliotecas digitais cooperação questionários bibliotecários entrevistas estudantes indicadores de C&T indicadores
ARTIGO 08 biblioteca digital; avaliação de bibliotecas digitais; metodologias de avaliação de bibliotecas digitais	bibliotecas digitais avaliação bibliografias	bibliotecas digitais categorias bibliografias levantamentos bibliotecas avaliação pesquisa
ARTIGO 09 bibliotecas digitais; acesso Aberto; publicações acadêmicas; repositórios institucionais; sistemas adaptativos complexos; conexidade; complexidade; software social	bibliotecas digitais acesso livre repositórios institucionais publicações economia da informação complexidade programas de computador	repositórios institucionais bibliotecas digitais economia da informação estudos de caso bibliotecas universitárias acesso livre coleções programas de computador inovação arquivos
ARTIGO 10 gestão da informação na área da saúde; informação em Engenharia biomédica; direitos autorais; biblioteca digital; usuário Virtual	bibliotecas digitais gestão da informação engenharia biomédica informação científica e tecnológica direito à informação direito autoral propriedade intelectual usuários	bibliotecas digitais direito autoral engenharia biomédica informação científica e tecnológica direito à informação tomada de decisões propriedade intelectual gestão da informação subsídios engenharias
ARTIGO 11 biblioteca digital; socialização da informação; e-book; núcleo temático da seca	bibliotecas digitais acesso livre livros eletrônicos digitalização	bibliotecas digitais ciências sociais aplicadas sociedade da informação acesso livre professores coleções livros digitalização acervos acesso
ARTIGO 12 governo eletrônico; arquitetura de sistemas de	bibliotecas digitais sistemas de informação informação governamental governo eletrônico	bibliotecas digitais informação governamental gestão da informação necessidades de informação

informação; integração de informações; gestão de C&T; bibliotecas digitais; plataforma lattes; rede scientia	arquitetura de informação gestão da informação dissertações e teses	governo eletrônico sistemas de informação inovação bibliotecas papel dissertações e teses
ARTIGO 13 arquitetura da informação; biblioteca digital; personalização	bibliotecas digitais arquitetura de informação disseminação da informação perfil do usuário usuários	arquitetura de informação bibliotecas digitais disseminação da informação acesso bibliotecas usuários
ARTIGO 14 bibliotecas digitais; bibliotecas universitárias; biblioteca eletrônica; biblioteca virtual	bibliotecas digitais bibliotecas virtuais bibliotecas universitárias redes de bibliotecas tecnologias da informação e comunicação organização do conhecimento	bibliotecas universitárias organização do conhecimento redes de bibliotecas tecnologias da informação e comunicação bibliotecas virtuais bases de dados bibliotecas digitais superposição estágios bibliotecários
ARTIGO 15 bilbioteca digital de teses e dissertações do ibct; usabilidade; avaliação analítica de usabilidade; avaliação empírica de usabilidade	bibliotecas digitais dissertações e teses usabilidade arquitetura de informação	bibliotecas digitais estudos de caso redes de telecomunicações recuperação da informação usabilidade eficiência pesquisa acesso dissertações e teses buscas
ARTIGO 16 biblioteca digital; arquitetura de biblioteca digital; projeto da biblioteca digital; normalização; arquitetura da informação; xml; Z39.50; digitalização Desenvolvimento de coleções; controle bibliográfico; catalogação; classificação metadados; ontologias; preservação digital; acesso; interface; interoperabilidade; referência digital; direitos Autorais; sustentabilidade; usuários de biblioteca; avaliação de biblioteca	bibliotecas digitais bibliografias fluxo da informação arquitetura de informação desenvolvimento de coleções controle bibliográfico digitalização catalogação preservação digital direito autoral acesso à informação	direito autoral arquitetura de informação desenvolvimento de coleções preservação digital bibliotecas digitais acesso à informação controle bibliográfico bibliografias XML catalogação
ARTIGO 17 biblioteca digital; horizon; impa; automação; preprint; metadados; xml; dublin core;	bibliotecas digitais pré-publicações dissertações e teses metadados Dublin Core XML	Dublin Core bibliotecas digitais aplicações de computador XML metadados pré-publicações

math net; impress; harvest		programas de computador programas de pós-graduação dissertações e teses bibliotecas
ARTIGO 18 bibliografia; biblioteca digital	bibliotecas digitais bibliografias grupos de discussão	grupos de discussão bibliotecas digitais treinamento manuais livros periódicos bibliografias pesquisa bibliotecas
ARTIGO 19 biblioteca digital; teses e dissertações eletrônicas; arquivos abertos; iniciativas de arquivos abertos; desenvolvimento de sistemas de informação; padrões de metadados; bdtd	bibliotecas digitais acesso livre dissertações e teses padrões de metadados metadados sistemas de informação	padrões de metadados acesso livre bibliotecas digitais sistemas de informação metadados normas e protocolos livros planejamento OAIS arquivos
ARTIGO 20 biblioteca digital; educação a distância; repositório digital; tecnologias da informação e comunicação; repositório digital EAD	bibliotecas digitais repositórios digitais ensino a distância tecnologias da informação e comunicação acesso remoto recursos de informação	acesso remoto repositórios digitais tecnologias da informação e comunicação recursos de informação bibliotecas digitais terminologia ensino a distância comunicação acesso pesquisa
ARTIGO 21 indexação automática; sistema de recuperação da informação; teatro; indexação e resumos; pesquisa; bibliotecas digitais	bibliotecas digitais recuperação da informação sistemas de recuperação da informação indexação automática indexação resumos	indexação automática bibliotecas digitais sistemas de recuperação da informação entrada de dados bibliotecas centrais indexação recuperação da informação entradas resumos normas e protocolos
ARTIGO 22 biblioteca digital; universidade do estado do rio de janeiro; teses e dissertações; produção acadêmica; bdtd	bibliotecas digitais universidades dissertações e teses disseminação da informação comunidades científicas	comunidades científicas bibliotecas digitais disseminação da informação avaliação universidades papel educação bibliotecas dissertações e teses
ARTIGO 23 ciência da informação; arquitetura da informação; biblioteca digital; teses e dissertações; informação;	bibliotecas digitais arquitetura de informação dissertações e teses usos da informação ciência da informação	arquitetura de informação usos da informação bibliotecas digitais ciência da informação eficiência acesso

tecnologia		pesquisa bibliotecas dissertações e teses
ARTIGO 24 biblioteca digital; direitos Autorais; biblioteca virtual	bibliotecas digitais bibliotecas virtuais direito autoral tecnologias da informação e comunicação Internet	bibliotecas virtuais bibliotecas digitais sociedade da informação tecnologias da informação e comunicação direito autoral Internet BD World Wide Web paradigmas livros
ARTIGO 25 biblioteca digital; competência informacional; educação a distância; gpcin	bibliotecas digitais competência em informação ensino a distância seleção de documentos acesso ao documento	competência em informação seleção de documentos acesso ao documento bibliotecas digitais estudantes ensino a distância acervos física bibliotecas educação
ARTIGO 26 biblioteca digital; biblioteca digital universal; acesso universal a informação	bibliotecas digitais acesso universal acesso à informação	acesso universal unidades de informação bibliotecas digitais acesso à informação cenários acervos história papel bibliotecas acesso
ARTIGO 27 biblioteca digital; estudo de usuário; recuperação de informação; tecnologia agropecuária	bibliotecas digitais perfil do usuário estudos de usuários recuperação da informação	perfil do usuário estudos de usuários bibliotecas digitais recuperação da informação questionários terminologia aplicativos empresas pesquisa navegação
ARTIGO 28 biblioteca em tempo real; biblioteca in loco; biblioteca híbrida; organização da informação; cibercultura; biblioteca virtual; biblioteca digital	bibliotecas digitais bibliotecas virtuais bibliotecas híbridas documentos eletrônicos organização do conhecimento	bibliotecas híbridas documentos eletrônicos bibliotecas virtuais organização do conhecimento transferência da informação fluxo da informação acesso à informação bibliotecas digitais acessibilidade bibliotecas
ARTIGO 29 biblioteca virtual; biblioteca digital; pesquisa científica no Brasil; comunicação	bibliotecas digitais bibliotecas virtuais comunicação científica Internet usuários	bibliotecas virtuais bibliotecas digitais comunicação científica recursos de informação Internet questionários

científica; internet		comunicação usuários pesquisadores papel
ARTIGO 30 biblioteca digital; acesso universal a informação; conteúdos digitais interativos	bibliotecas digitais acesso universal acesso à informação cenários	acesso universal bibliotecas digitais cenários subsídios acesso bibliotecas
ARTIGO 31 biblioteca digital; biblioteca digital contextualizada	bibliotecas digitais World Wide Web Internet	bibliotecas digitais World Wide Web Internet conceitos bibliotecas
ARTIGO 32 Bases de dados; Estratégia de busca; Linguagem controlada; Linguagem natural. Recuperação da informação; Artigo de revisão	bibliotecas digitais bibliotecas virtuais anais de congressos comunicação científica produtividade científica ciência da informação	bibliotecas universitárias bibliotecas virtuais bibliotecas digitais ciência da informação periódicos biblioteconomia comunicação autoria bibliotecas
ARTIGO 33 Biblioteca digital; Biblioteca virtual; Produção científica; Produção bibliográfica; Periódicos	bibliotecas digitais bibliotecas virtuais produtividade científica artigos de periódico periódicos científicos	referências bibliográficas bibliotecas virtuais bibliotecas digitais tipos de documento desenvolvimento de coleções unidades de informação produtividade de autor artigos de periódico profissionais de informação planejamento
ARTIGO 34 tecnologia da informação; categorização; biblioteca digital; mineração de texto; documentos digitais	bibliotecas digitais documentos eletrônicos categorização automática de textos classificação automática mineração de textos	documentos eletrônicos estudos de caso mineração de textos tecnologias da informação e comunicação bibliotecas digitais física algoritmos categorias World Wide Web catalogação
ARTIGO 35 bibliotecas digitais; bibliotecas nacionais; digitalização; disseminação da informação; colóquio bibliotecas digitais brasil frança alemanha	bibliotecas digitais bibliotecas nacionais acervos digitalização disseminação da informação	bibliotecas nacionais bibliotecas digitais digitalização bibliotecas acervos
ARTIGO 36 biblioteca digital; competência profissional;	bibliotecas digitais preservação digital competências profissionais profissionais de informação	preservação digital bibliotecas digitais gestão de bibliotecas serviços de informação

gestão de biblioteca; preservação digital; profissional da informação	gestão de bibliotecas	competências profissionais profissionais de informação bibliotecas gestão papel gestores
ARTIGO 37 acesso a informação; biblioteca digital; bibliotecário; cooperação bibliotecária; digitalização; google; internet	bibliotecas digitais organização do conhecimento acesso à informação digitalização cooperação entre bibliotecas bibliotecários mecanismos de busca Internet	cooperação entre bibliotecas mecanismos de busca organização do conhecimento bibliotecas digitais acesso à informação custos cooperação bibliotecários Internet digitalização
ARTIGO 38 digitalização; obras raras; biblioteca digital; preservação	bibliotecas digitais digitalização obras raras conservação de documentos acesso à informação	obras raras bibliotecas digitais acesso à informação acervos digitalização bibliotecas acesso
ARTIGO 39 digitalização; preservação digital; biblioteca acadêmica; biblioteca digital	bibliotecas digitais bibliotecas universitárias digitalização preservação digital conservação de documentos	preservação digital bibliotecas digitais bibliotecas universitárias equipamentos de computador digitalização programas de computador acervos acesso universidades bibliotecas
ARTIGO 40 bibliotecas digitais; website; instrumentos de pesquisa; bibliotecas universitárias	bibliotecas digitais bibliotecas universitárias programas livres programas de computador páginas da web sítios web	bibliotecas universitárias programas livres bibliotecas digitais catálogos livros educação pesquisa bibliotecas programas de computador
ARTIGO 41 biblioteca digital; cloud computing; ciência da informação	bibliotecas digitais redes de bibliotecas ciência da informação documentos eletrônicos registros bibliográficos publicações oficiais	bibliotecas digitais redes de bibliotecas documentos eletrônicos registros bibliográficos publicações oficiais ciência da informação probabilidade e estatística registros publicações bibliotecas
ARTIGO 42 5S (societies = sociedades, scenarios = cenários, spaces = espaços, structures = structures, streams = correntes); Currículos; DL	bibliotecas digitais repositórios institucionais acesso livre dissertações e teses normas e protocolos	bibliotecas digitais repositórios institucionais acesso livre OAIS normas e protocolos cenários arquivos estudantes

(digital library = bibliotecas digitais); ETD (electronic thesis or dissertations = teses ou dissertações eletrônicas); NDLTD (networked digital library of theses and dissertations = biblioteca digital em rede de teses e dissertações); Acesso aberto; OAI (open archives initiative = iniciativa dos arquivos abertos); Padrões		pesquisa bibliotecas
ARTIGO 43 gestão da informação; gestão do conhecimento; biblioteca digital; indicadores infraero; ministério da saúde	bibliotecas digitais gestão da informação gestão do conhecimento indicadores indicadores de C&T	bibliotecas digitais estudos de caso gestão do conhecimento gestão da informação indicadores de C&T indicadores empresas BD pesquisa gestão
ARTIGO 44 ensino a distância; informação digital; bibliotecas digitais	bibliotecas digitais ensino a distância documentos eletrônicos direito autoral	documentos eletrônicos bibliotecas digitais direito autoral ensino a distância estudantes manutenção acervos paradigmas gestão papel
ARTIGO 45 bibliotecas digitais; publicações eletrônicas; arquivos abertos; interoperabilidade; metadados; padrões; tecnologia da informação; informação em ciência e tecnologia; comunicação científica; acesso a informação	bibliotecas digitais publicações eletrônicas acesso livre interoperabilidade metadados normas e protocolos comunicação científica acesso à informação tecnologias da informação e comunicação portais	publicações eletrônicas comunicação científica artigos de periódico tecnologias da informação e comunicação recursos de informação acesso à informação bibliotecas digitais sistemas de informação necessidades de informação portais
ARTIGO 46 biblioteca digital; educação; usabilidade; interatividade; estudo de usuário	bibliotecas digitais educação buscas de informação usabilidade interação homem-computador estudos de usuários	estudos de usuários buscas de informação redes de telecomunicações revisões de literatura bibliotecas digitais ciência da informação Internet tecnologias da informação e comunicação usabilidade psicologia

ARTIGO 47 metadados; biblioteca digital; preservação digital; METS; Metadata Encoding & Transmission Standard	bibliotecas digitais metadados preservação digital Metadata Encoding and Transmission Standard padrões de metadados XML	preservação digital objetos digitais bibliotecas digitais normas e protocolos XML Metadata Encoding and Transmission Standard metadados acervos bibliotecas
ARTIGO 48 biblioteca; biblioteca digital; livro eletrônico; computador; internet; tecnologia; educação	bibliotecas digitais livros eletrônicos bibliotecas computadores Internet educação ensino a distância tecnologias da informação e comunicação	livros eletrônicos bibliotecas digitais multimídia ensino a distância lógica livros registros computadores Internet educação
ARTIGO 49 modelos conceituais; bibliotecas digitais; sistema de biblioteca digital	bibliotecas digitais modelos cognitivos redes de bibliotecas organização do conhecimento classificação facetada	classificação facetada modelos cognitivos redes de bibliotecas organização do conhecimento bibliotecas digitais recuperação da informação classificação mapas acervos termos
ARTIGO 50 informação científica; arquivos abertos; acesso livre; bibliotecas digitais; repositórios digitais; bdtd; iniciativa dos arquivos abertos; acesso a informação científica; crise dos periódicos	bibliotecas digitais repositórios digitais acesso livre informação científica e tecnológica comunicação científica	bibliotecas digitais periódicos científicos tecnologias da informação e comunicação repositórios digitais informação científica e tecnológica comunicação científica acesso livre acesso à informação editores editoras
ARTIGO 51 Arquitetura da informação. Bibliotecas digitais. Website.	bibliotecas digitais arquitetura da informação sítios web ciência da informação biblioteconomia	arquitetura de informação bibliotecas digitais ciência da informação navegação buscas biblioteconomia bibliotecas
ARTIGO 52 biblioteca digital; biblioteca universitária; serviços de biblioteca para o ensino a distância	bibliotecas digitais bibliotecas universitárias acesso livre ensino a distância serviços de biblioteca universidades	bibliotecas universitárias ensino a distância serviços de biblioteca acesso livre bibliotecas digitais cooperação acesso universidades bibliotecas educação
ARTIGO 53	bibliotecas digitais informação científica e tecnológica	bibliotecas digitais informação científica e tecnológica

biblioteca digital; bibliotecário; e-science; ciberinfraestrutura; dados abertos; dilúvio de dados; quarto paradigma; ciência da informação; pesquisa colaborativa	dados científicos acesso livre profissionais de informação bibliotecários ciência da informação	ciência da informação cenários profissionais de informação paradigmas bibliotecários termos conceitos bibliotecas
ARTIGO 54 biblioteca digital; copyleft; documentos digitais; domínio público; preservação	bibliotecas digitais bibliotecas documentos eletrônicos recursos de informação domínio público preservação digital acessibilidade	documentos eletrônicos recursos de informação bibliotecas digitais acessibilidade coleções conceitos livros bibliotecas
ARTIGO 55 biblioteca digital; ensino de biblioteca digital; ensino de biblioteconomia	bibliotecas digitais ensino de biblioteconomia formação profissional biblioteconomia ensino superior	pesquisa exploratória ensino de biblioteconomia bibliotecas digitais professores formação profissional bibliografias pesquisa estudantes levantamentos biblioteconomia
ARTIGO 56 biblioteca digital; bibliotecas tradicionais; google books; tecnologia da informação; direitos Autorais; obras órfãs	bibliotecas digitais bibliotecas tecnologias da informação e comunicação empresas digitalização livros direito autoral	direito autoral bibliotecas digitais tecnologias da informação e comunicação digitalização empresas livros direito bibliotecas conceitos pesquisa
ARTIGO 57 biblioteca digital; obras raras; digitalização; preservação de documentos; preservação digital	bibliotecas digitais obras raras livros digitalização preservação digital preservação de documentos acesso remoto	acesso remoto preservação digital obras raras tipos de documento preservação de documentos bibliotecas digitais acervos digitalização livros coleções
ARTIGO 58 bibliotecas digitais; periódico eletrônico; legibilidade; usabilidade	bibliotecas digitais periódicos eletrônicos portais usabilidade arquitetura de informação	bibliotecas digitais periódicos eletrônicos comportamento do usuário revisões de literatura informação científica e tecnológica disseminação da informação acesso à informação ciência da informação usabilidade portais
ARTIGO 59 serviço de referência e	bibliotecas digitais serviços de referência eletrônica serviços de referência	revisões de literatura bibliotecários de referência unidades de informação

informação digital; serviço de referência; bibliotecário de referência digital; biblioteca digital	bibliotecários de referência	serviços de referência bibliotecas digitais ciência da informação sistemas de informação profissionais de informação eficiência sistemas de recuperação da informação
ARTIGO 60 biblioteca tradicional; biblioteca eletrônica; biblioteca eletrônica virtual; biblioteca polimídia; biblioteca interativa; biblioteca virtual; biblioteca de realidade virtual; biblioteca digital; biblioteca universal	bibliotecas digitais bibliotecas virtuais paradigmas bibliotecas	bibliotecas virtuais realidade virtual bibliotecas digitais paradigmas bibliotecas mudança
ARTIGO 61 recuperação de informação; metadados; bdtd	bibliotecas digitais dissertações e teses representação da informação metadados recuperação da informação indexação Código de Catalogação Anglo-Americano	representação da informação organização do conhecimento tecnologias da informação e comunicação bibliotecas digitais acesso livre recuperação da informação física metadados Código de Catalogação Anglo-Americano indexação
ARTIGO 62 repositório institucional; Política institucional de acesso aberto; repositório da produção científica do Cruesp; biblioteca digital da produção intelectual da USP; portal de Revistas da USP; formulação de políticas; portais de acesso a revistas científicas; acesso aberto	repositórios institucionais bibliotecas digitais acesso livre periódicos científicos portais cultura organizacional políticas de informação informação científica e tecnológica	cultura organizacional periódicos científicos bibliotecas digitais projetos de pesquisa repositórios institucionais acesso livre portais bibliotecários cenários estudantes
ARTIGO 63 direitos Autorais; direito a informação; biblioteca digital paulo freire; democratização da informação	bibliotecas digitais documentos eletrônicos direito autoral digitalização direito à informação acesso à informação	documentos eletrônicos bibliotecas digitais direito autoral direito à informação acesso à informação acervos digitalização Internet acesso pesquisa
ARTIGO 64 biblioteca digital; repositório	bibliotecas digitais repositórios institucionais acesso livre	repositórios institucionais análise comparativa acesso livre

institucional; monitorização; política mandatória	publicações bases de dados políticas de informação monitoramento	bibliotecas digitais bases de dados probabilidade e estatística normalização gráficos acesso publicações
ARTIGO 65 bdtd; ciências humanas e sociais; disseminação do conhecimento; acesso livre	bibliotecas digitais dissertações e teses acesso livre programas de pós-graduação ciências humanas disseminação da informação	ciências humanas acesso livre bibliotecas digitais manutenção livros programas de pós-graduação acesso pesquisa dissertações e teses usuários
ARTIGO 66 biblioteca digital federada; OAI-PMH; metadados; bibliometria	bibliotecas digitais dissertações e teses periódicos científicos acesso livre bibliometria metadados OAI-PMH	comunidades científicas periódicos científicos bibliotecas digitais acesso livre OAIS bibliometria metadados normas e protocolos livros periódicos
ARTIGO 67 Tecnologias da Informação e Comunicação; Educação; Biblioteca Digital de Teses e Dissertações; IBICT I	bibliotecas digitais dissertações e teses ensino a distância educação programas de pós-graduação tecnologias da informação e comunicação	pesquisa exploratória ensino a distância tecnologias da informação e comunicação bibliotecas digitais programas de pós-graduação dissertações e teses comunicação bibliotecas educação pesquisa
ARTIGO 68 biblioteca digital multilíngue; metadados; base de dados multilíngues	bibliotecas digitais livros eletrônicos metadados bases de dados barreiras linguísticas linguagens documentárias	estudos de caso livros eletrônicos bibliotecas digitais metadados funcionalidade livros Internet autoria dados usuários
ARTIGO 69 biblioteca digital geográfica; geoprocessamento; sistema de informação geográfica; metadados; Geo-ontologias; biblioteca digital geográfica distribuída	bibliotecas digitais geoinformação representação da informação padrões de metadados ontologias sistemas de informação geográfica sistemas de informação	objetos digitais geoinformação acesso livre padrões de metadados redes de telecomunicações teorias na ciência da informação representação da informação tecnologias da informação e comunicação bibliotecas digitais ciência da informação

ARTIGO 70 biblioteca escolar; centro de recursos para o ensino e a aprendizagem; gestão de conteúdos digitais educativos; biblioteca digital educativa	bibliotecas digitais bibliotecas escolares educação sociedade da informação gestão de conteúdos na web	bibliotecas digitais bibliotecas escolares estudos de caso organização do conhecimento sociedade da informação estados da arte artes gestão educação bibliotecas
ARTIGO 71 preservação digital; metadados; repositórios digitais; modelo de preservação OAIS; biblioteca digital	bibliotecas digitais repositórios digitais preservação digital metadados OAIS preservação de documentos objetos digitais documentos eletrônicos	objetos digitais repositórios digitais documentos eletrônicos preservação digital bibliotecas digitais preservação de documentos metadados OAIS bibliotecas pesquisa
ARTIGO 72 acesso livre; informação científica; biblioteca digital; repositórios digitais	bibliotecas digitais repositórios digitais acesso livre informação científica e tecnológica recursos de informação coleções especiais	repositórios digitais informação científica e tecnológica bibliotecas digitais acesso livre tecnologias da informação e comunicação recursos de informação similaridade livros estudantes história
ARTIGO 73 ontologias; epistemologia; representação do conhecimento; metadados; conceitos; ciência da informação; bibliotecas digitais; arquivos digitais; hermenêutica	bibliotecas digitais representação do conhecimento ontologias ciência da informação epistemologia hermenêutica	representação do conhecimento acesso à informação bibliotecas digitais ciência da informação sistemas de informação transdisciplinaridade hermenêutica títulos de documentos metadados epistemologia
ARTIGO 74 serviços de referência virtual; biblioteca digital; tecnologia da informação; referência digital; bibliotecário de referência; correio eletrônico	bibliotecas digitais serviços de referência serviços de referência eletrônica bibliotecários de referência tecnologias da informação e comunicação	correio eletrônico serviços de referência bibliotecas digitais bibliotecários de referência tecnologias da informação e comunicação bibliotecários normas e protocolos bibliotecas
ARTIGO 75 biblioteca digital; taxonomia facetada; interface de busca; acesso à informação	bibliotecas digitais acesso à informação taxonomias classificação facetada sistemas de recuperação da informação recuperação da informação buscas de informação	bibliotecas digitais sistemas de recuperação da informação acesso à informação recuperação da informação taxonomias classificação dados buscas bibliotecas

		acesso
ARTIGO 76 bibliotecas digitais; sistemas de informação; biblioteca digital universitária	bibliotecas digitais sistemas de informação bibliotecas universitárias serviços de biblioteca dissertações e teses propriedade intelectual universidades	serviços de biblioteca bibliotecas digitais propriedade intelectual sistemas de informação manutenção gestão bibliotecas pesquisa universidades dissertações e teses
ARTIGO 77 teses e dissertações eletrônicas; teses e dissertações; universidade de los angeles; venezuela; software; repositórios institucionais; adaptação; software livre; biblioteca digital	bibliotecas digitais repositórios institucionais dissertações e teses programas livres programas de computador universidades	programas livres repositórios institucionais bibliotecas digitais programas de computador livros funcionalidade teste publicações bibliotecas universidades
ARTIGO 78 biblioteca digital; gestão da informação e do conhecimento; transferência da informação e do conhecimento; pequenas e médias empresas	bibliotecas digitais transferência da informação acesso à informação pequenas e médias empresas gestão da informação gestão do conhecimento universidades empresas	bibliotecas digitais ciências sociais aplicadas publicações seriadas artigos de periódico literatura cinzenta pequenas e médias empresas transferência da informação acesso à informação ciência da informação empresas
ARTIGO 79 biblioteca digital; arquivos abertos; informetria; bibliometria; banco de teses; intercâmbio e compartilhamento de informações; gestão do conhecimento	bibliotecas digitais gestão do conhecimento informetria bibliometria dissertações e teses	bibliotecas digitais engenharia de produção documentos eletrônicos gestão do conhecimento bibliometria informetria professores engenharias arquivos estudantes
ARTIGO 80 usabilidade; comportamento de busca por informação; biblioteca digital	bibliotecas digitais usuários comportamento do usuário buscas de informação sistemas de informação usabilidade	usuários novatos buscas de informação bibliotecas digitais ciência da computação ciência da informação questionários sistemas de informação teste livros usabilidade
ARTIGO 81 imagens em medicina; biblioteca virtual; biblioteca digital; bancos de imagens; preservação da informação	bibliotecas digitais bibliotecas virtuais imagens medicina bases de dados de imagens preservação digital linguagens documentárias metadados	linguagens documentárias bibliotecas virtuais bibliotecas digitais metadados acervos imagens medicina bibliotecas

		gestão educação
ARTIGO 82 biblioteca digital; intercâmbio de informação	bibliotecas digitais economia da informação transferência da informação serviços de biblioteca competências profissionais	economia da informação tecnologias da informação e comunicação serviços de biblioteca bibliotecas digitais transferência da informação economia mudança bibliotecas

ANEXO A – TABELA COM AS REFERÊNCIAS DOS 60 ARTIGOS DO *CORPUS* A-SOUZA

Artigo 1 - SANTOS, P. L. V. A. C.; SANTANA, R. C. G. Transferência da informação: análise para valoração de unidades de conhecimento. **DataGramaZero**, [*s.l.*], v. 3, n. 2, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5349. Acesso em: 16 out. 2021.
Artigo 2 - MUELLER, S. P. M. Popularização do conhecimento científico. **DataGramaZero**, [*s.l.*], v. 3, n. 2, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5354. Acesso em: 16 out. 2021.
Artigo 3 - CASTRO, A. L. S. O valor da informação: um desafio permanente. **DataGramaZero**, [*s.l.*], v. 3, n. 3, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5365. Acesso em: 16 out. 2021.
Artigo 4 - LAGE, Márcia Basílio; CAFÉ, Ligia. Auto-arquivamento: uma opção inovadora para a produção científica. **DataGramaZero,** [*s.l.*], v. 3, n. 3, 2001. Disponível em: https://ridi.ibict.br/bitstream/123456789/280/1/CAFE2002.pdf. Acesso em: 16 out. 2021.
Artigo 5 - BURNHAM, T. F. Análise contrastiva: memória da construção de uma metodologia para investigar a tradução de conhecimento científico em conhecimento público. **DataGramaZero**, [*s.l.*], v. 3, n. 3, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/6809. Acesso em: 16 out. 2021.
Artigo 6 - LEVACOV, M.; VANTI, N.; ZANCAN, J. C.; MENDES, M. L. G. O tesauro eletrônico do mundo do trabalho: produto de um esforço interdisciplinar. **DataGramaZero**, [*s.l.*], v. 3, n. 4, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/6816. Acesso em: 16 out. 2021.
Artigo 7 - VALENTIM, M. L. P. Inteligência competitiva em organizações: dado, informação e conhecimento. **DataGramaZero**, [*s.l.*], v. 3, n. 4, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/3837. Acesso em: 16 out. 2021.
Artigo 8 - MIRANDA, A. B.; SIMEÃO, E. L. M. S. A conceituação de massa documental e o ciclo de interação entre tecnologia e o registro do conhecimento. **DataGramaZero**, [*s.l.*], v. 3, n. 4, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/3853. Acesso em: 16 out. 2021.
Artigo 9 - STAREC, Claudio. Informação e universidade: os pecados informacionais e barreiras na comunicação da informação para a tomada de decisão na universidade. **Datagramazero**, [*s.l.*], v. 3, n. 4, p. 1-11, 2002. Disponível em: http://afro.culturadigital.br/wp-content/uploads/2017/10/Artigo-09-1.pdf. Acesso em: 16 out. 2021.
Artigo 10 - PORCARO, R. M. Implicações da ´nova economia´ para a mensuração estatística: desajustes conceituais e metodológicos. **DataGramaZero**, [*s.l.*], v. 3, n. 4, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5382. Acesso em: 16 out. 2021.
Artigo 11 - PATERNOSTRO, Luiz Carlos Brito. Por uma nova Ciência da Informação: ensino, pesquisa e formação. **DataGramaZero**, [*s.l.*], v. 3, n. 5, 2002. Disponível em: http://afro.culturadigital.br/wp-content/uploads/2017/10/Artigo-11.pdf. Acesso em: 16 out. 2021.
Artigo 12 - DIAS, Eduardo Wense. Ensino e pesquisa em ciência da informação. **DataGramaZero**, Rio de Janeiro, v. 3, n. 5, 2002. Disponível em: http://afro.culturadigital.br/wp-content/uploads/2017/10/Artigo-12-1.pdf. Acesso em: 16 out. 2021.
Artigo 13 - CARVALHO, K. O profissional da informação: o humano multifacetado. **DataGramaZero**, [*s.l.*], v. 3, n. 5, 2002. Disponível em:

http://hdl.handle.net/20.500.11959/brapci/5395. Acesso em: 16 out. 2021.
Artigo 14 - TARAPANOFF, Kira Maria Antonia; SUAIDEN, Emir José; OLIVEIRA, Cecilia Leite. Funções sociais e oportunidades para profissionais da informação. **DataGramaZero,** [*s.l.*], v. 3, n. 5, 2002. Disponível em: https://ridi.ibict.br/handle/123456789/256. Acesso em: 16 out. 2021.
Artigo 15 - RODRIGUES, M. E. F. Relação ensino-pesquisa: em discussão a formação do profissional da informação. **DataGramaZero**, [*s.l.*], v. 3, n. 5, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5410. Acesso em: 14 mar. 2021.
Artigo 16 - CARDOSO, A. M. P. Educação para a informação: desafios contemporâneos para a ciência da informação. **DataGramaZero**, [*s.l.*], v. 3, n. 5, 2002. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/5414>. Acesso em: 16 out. 2021.
Artigo 17 - TARGINO, M. D. G. Novas tecnologias e produção científica: uma relação de causa e efeito ou uma relação de muitos efeitos? **DataGramaZero**, [*s.l.*], v. 3, n. 6, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5421. Acesso em: 16 out. 2021.
Artigo 18 - DAGNINO, R. P. Enfoques sobre a relação ciência, tecnologia e sociedade: neutralidade e determinismo. **DataGramaZero**, [*s.l.*], v. 3, n. 6, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5429. Acesso em: 16 out. 2021.
Artigo 19 - BARBOSA, Ricardo Rodrigues et al. Inteligência empresarial: uma avaliação de fontes de informação sobre o ambiente organizacional externo. **DataGramaZero**, [*s.l.*], v. 3, n. 6, 2002. Disponível em: https://www.brapci.inf.br/_repositorio/2010/01/pdf_8c57e423fa_0007498.pdf. Acesso em: 16 out. 2021.
Artigo 20 - SMIT, J. W.; DIAS, E. J. W.; SOUZA, R. F. Contribuição da pós-graduação para a ciência da informação no brasil: uma visão. **DataGramaZero**, [*s.l.*], v. 3, n. 6, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5440. Acesso em: 16 out. 2021.
Artigo 21 - DUMONT, L. M. M. Os múltiplos aspectos e interfaces da leitura. **DataGramaZero,** [*s.l.*], v. 3, n. 6, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5446. Acesso em: 16 out. 2021.
Artigo 22 - SANTOS, N. B. D. A informação e o paradigma holográfico: a utopia de vannevar bush. **DataGramaZero**, [*s.l.*], v. 3, n. 6, 2002. Disponível em: http://hdl.handle.net/20.500.11959/brapci/6828. Acesso em: 16 out. 2021.
Artigo 23 - THIESEN, I. Informação, memória e espaço prisional no rio de janeiro. **DataGramaZero**, [*s.l.*], v. 4, n. 1, 2003. Disponível em: http://hdl.handle.net/20.500.11959/brapci/6779. Acesso em: 14 mar. 2021.
Artigo 24 - GÓNZALEZ DE GÓMEZ, Maria Nélida. O contrato social da pesquisa: em busca de uma nova equação entre a autonomia epistêmica e autonomia política. **DataGramaZero**, [*s.l.*], v. 4, n. 1, 2003. Disponível em: https://ridi.ibict.br/handle/123456789/122. Acesso em: 16 out. 2021.
Artigo 25 - MUELLER, Suzana Pinheiro Machado; SANTANA, Maria Gorette Henrique. A Ciência da Informação no CNPq: fomento à formação de recursos humanos e à pesquisa entre 1994-2002. **DataGramaZero,** [*s.l.*], v. 4, n. 1, 2003. Disponível em: https://repositorio.unb.br/handle/10482/963. Acesso em: 16 out. 2021.
Artigo 26 - BARRETO, A. A. Políticas de monitoramento da informação por compressão semântica dos seus estoques. **DataGramaZero**, [*s.l.*], v. 4, n. 2, 2003. Disponível em: http://hdl.handle.net/20.500.11959/brapci/4032. Acesso em: 16 out. 2021.
Artigo 27 - OLINTO, G. Bolsas de pesquisador do cnpq: informações sobre política de c&t a partir da base que contém os dados cadastrais dos bolsistas. **DataGramaZero**, [*s.l.*], v. 4, n. 2, 2003. Disponível em: http://hdl.handle.net/20.500.11959/brapci/6787. Acesso em: 16 out. 2021.
Artigo 28 - PACHECO, R. C. D. S.; KERN, V. M. Arquitetura conceitual e resultados da

integração de sistemas de informação e gestão da ciência e tecnologia. **DataGramaZero**, [*s.l.*], v. 4, n. 2, 2003. Disponível em: http://hdl.handle.net/20.500.11959/brapci/3877. Acesso em: 16 out. 2021..
Artigo 29 - MARCONDES, C. H.; JARDIM, J. M. Políticas de informação governamental: a construção de governo eletrônico na administração federal do brasil. **DataGramaZero**, [*s.l.*], v. 4, n. 2, 2003. Disponível em: http://hdl.handle.net/20.500.11959/brapci/3900. Acesso em: 16 out. 2021.
Artigo 30 - DIAS, Guilherme Ataíde. Avaliação do acesso a periódicos eletrônicos na web pela análise do arquivo de log de acesso. **Ciência da Informação**, Brasília, v. 31, n. 1, p. 7-12, jan. 2002. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000100002&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 31 - GONZALEZ DE GOMEZ, Maria Nélida. Novos cenários políticos para a informação. **Ciência da Informação**, Brasília, v. 31, n. 1, p. 27-40, jan. 2002. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000100004&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 32 - LOPES, Ilza Leite. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. **Ciência da Informação**, Brasília , v. 31, n. 1, p. 41-52, jan. 2002. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000100005&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 33 - OHIRA, M. L. B.; PRADO, N. S. Bibliotecas virtuais e digitais: análise de artigos de periódicos brasileiros (1995/2000). **Ciência da Informação**, Brasília, v. 31, n. 1, 2002. DOI: 10.18225/ci.inf..v31i1.978 Acesso em: 16 out. 2021.
Artigo 34 - PRYSTHON, Cecília; SCHMIDT, Susana. Experiência do Leaal/UFPE na produção e transferência de tecnologia. **Ciência da informação**, Brasília, v. 31, p. 84-90, 2002. Disponível em: https://www.scielo.br/j/ci/a/fKk7D8w4ZMhzTxkXW74JgFN/abstract/?lang=pt. Acesso em: 16 out. 2021.
Artigo 35 - ALMEIDA, M. B. Uma introdução ao xml, sua utilização na internet e alguns conceitos complementares. **Ciência da Informação**, v. 31, n. 2, 2002. DOI: 10.18225/ci.inf..v31i2.955 Acesso em: 14 mar. 2021. Disponível em: http://revista.ibict.br/ciinf/article/view/955. Acesso em: 16 out. 2021.
Artigo 36 - ALVARADO-URBIZAGASTEGUI, R. A lei de lotka: modelo lagrangiano de poisson aplicado a produtividade de autores. **Perspectivas em Ciência da Informação**, [*s.l.*], v. 8, n. 2, 2003. Disponível em: http://hdl.handle.net/20.500.11959/brapci/35679. Acesso em: 16 out. 2021.
Artigo 37 - CENDÓN, B. V. Bases de dados de informação para negócios no brasil. **Ciência da Informação**, [*s.l.*], v. 32, n. 2, 2003. Disponível em: https://www.scielo.br/j/ci/a/WvMSbdCC9zMxB9QnQQR4fsh/abstract/?lang=pt. Acesso em: 16 out. 2021.
Artigo 38 - GARCEZ, E. M. S.; RADOS, G. J. V. Biblioteca híbrida: um novo enfoque no suporte à educação a distância. **Ciência da Informação**, [*s.l.*], v. 31, n. 2, 2002. DOI: 10.18225/ci.inf..v31i2.959 Acesso em: 16 out. 2021.
Artigo 39 - LOPES, Ilza Leite. Estratégia de busca na recuperação da informação: revisão da literatura. **Ciência Informação**, Brasília, v. 31, n. 2, p. 60-71, aug. 2002. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000200007&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 40 - MARCHIORI, Patricia Zeni. A ciência e a gestão da informação: compatibilidades no espaço profissional. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 72-79, aug. 2002. Disponível em:

http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000200008&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 41 - AGUILAR-POBLACIóN, D.; NORONHA, D. P. Produção das literaturas "branca" e "cinzenta" pelos docentes/doutores dos programas de pós-graduação em ciência da informação no brasil. **Ciência da Informação**, [*s.l.*], v. 31, n. 2, 2002. DOI: 10.18225/ci.inf..v31i2.965. Acesso em: 16 out. 2021.
Artigo 42 - REZENDE, Y. Informação para negócios: os novos agentes do conhecimento e a gestão do capital intelectual. **Ciência da Informação**, [*s.l.*], v. 31, n. 1, 2002. DOI: 10.18225/ci.inf..v31i1.979. Acesso em: 16 out. 2021.
Artigo 43 - SILVA, Janete Fernandes; FERREIRA, Marta Araújo Tavares; BORGES, Mônica Erichsen Nassif. Análise metodológica dos estudos de necessidades de informação sobre setores industriais brasileiros: proposições. **Ciência da informação**, [*s.l.*], v. 31, p. 129-141, 2002. Disponível em: https://www.scielo.br/j/ci/a/qYwrcFVNygWkxxfbxqyszSk/?format=pdf&lang=pt. Acesso em: 16 out. 2021.
Artigo 44 - SILVA, Sergio Luis da. Informação e competitividade: a contextualização da gestão do conhecimento nos processos organizacionais. **Ciência da Informação**, Brasília, v. 31, n. 2, p. 142-151, Aug. 2002. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000200015&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 45 - VANTI, N. A. P. Da bibliometria à webometria: uma exploração conceitual dos mecanismos utilizados para medir o registro da informação e a difusão do conhecimento. **Ciência da Informação**, [*s.l.*], v. 31, n. 2, 2002. DOI: 10.18225/ci.inf..v31i2.970. Acesso em: 16 out. 2021.
Artigo 46 - BORGES, P. C. R. Métodos quantitativos de apoio à bibliometria:a pesquisa operacional pode ser uma alternativa? **Ciência da Informação**, [*s.l.*], v. 31, n. 3, 2002. DOI: 10.18225/ci.inf..v31i3.943 Acesso em: 16 out. 2021.
Artigo 47 - DIAS, Guilherme Ataíde. Periódicos eletrônicos: considerações relativas à aceitação deste recurso pelos usuários. **Ciência da Informação**, Brasília, v. 31, n. 3, p. 18-25, Sept. 2002. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000300002&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 48 - COHEN, M. F. Alguns aspectos do uso da informação na economia da informação. **Ciência da Informação**, [*s.l.*], v. 31, n. 3, 2002. DOI: 10.18225/ci.inf.v31i3.945 Acesso em: 16 out. 2021.
Artigo 49 - ORTIZ, L. C.; ORTIZ, W. A.; SILVA, S. L. Ferramentas alternativas para monitoramento e mapeamento automatizado do conhecimento. **Ciência da Informação**, [*s.l.*], v. 31, n. 3, 2002. DOI: 10.18225/ci.inf..v31i3.949 Acesso em: 16 out. 2021.
Artigo 50 - SILVA, E. L.; CUNHA, M. F. V. A formação profissional no século xxi: desafios e dilemas. **Ciência da Informação**, [*s.l.*], v. 31, n. 3, 2002. DOI: 10.18225/ci.inf..v31i3.950 Acesso em: 16 out. 2021.
Artigo 51 - TORRES, E. F.; MAZZONI, A. A.; ALVES, J. B. M. A acessibilidade à informação no espaço digital. **Ciência da Informação**, [*s.l.*], v. 31, n. 3, 2002. DOI: 10.18225/ci.inf..v31i3.951 Acesso em: 16 out. 2021.
Artigo 52 - FREIRE, I. M.; NATHANHSON, B. M.; TAVARES, C.; SANTO, C. E. Estudos de usuários: o padrão que une três abordagens. **Ciência da Informação**, [*s.l.*], v. 31, n. 3, 2002. DOI: 10.18225/ci.inf..v31i3.953 Acesso em: 16 out. 2021.
Artigo 53 - AIRES, Rachel Virgínia Xavier; ALUISIO, Sandra Maria. Como incrementar a qualidade dos resultados das máquinas de busca: da análise de logs à interação em português **Ciência da Informação**, Brasília, v. 32, n. 1, p. 05-16, abr. 2003. Disponível

em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652003000100001&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 54 - DUDZIAK, E. A. Information literacy: princípios, filosofia e prática. **Ciência da Informação**, v. 32, n. 1, 2003. DOI: 10.18225/ci.inf..v32i1.1016 Acesso em: 16 out. 2021.
Artigo 55 - FERREIRA, D. T. Profissional da informação: perfil de habilidades demandadas pelo mercado de trabalho. **Ciência da Informação**, [*s.l.*], v. 32, n. 1, 2003. DOI: 10.18225/ci.inf..v32i1.1018 Acesso em: 16 out. 2021.
Artigo 56 - FREIRE, I. M. O olhar da consciência possível sobre o campo científico. **Ciência da Informação**, [*s.l.*], v. 32, n. 1, 2003. DOI: 10.18225/ci.inf..v32i1.1019. Acesso em: 16 out. 2021.
Artigo 57 - GÓMEZ, M. N. G. As relações entre ciência, estado e sociedade: um domínio de visibilidade para as questões da informação. **Ciência da Informação**, [*s.l.*], v. 32, n. 1, 2003. DOI: 10.18225/ci.inf..v32i1.1020 Acesso em: 16 out. 2021.
Artigo 58 - LIMA, Gercina Ângela Borém. Interfaces entre a ciência da informação e a ciência cognitiva. **Ciência da Informação,** Brasília, v. 32, n. 1, p. 77-87, abr. 2003. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652003000100008&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 59 - MOSTAFA, S. P.; MáXIMO, L. F. A produção científica da anped e da intercom no gt da educação e comunicação. **Ciência da Informação**, [*s.l.*], v. 32, n. 1, 2003. DOI: 10.18225/ci.inf..v32i1.1023 Acesso em: 16 out. 2021.
Artigo 60 - SILVA, H. P. Inteligência competitiva na internet: um processo otimizado por agentes inteligentes. **Ciência da Informação**, [*s.l.*], v. 32, n. 1, 2003. DOI: 10.18225/ci.inf..v32i1.1025. Acesso em: 16 out. 2021.

ANEXO B – TABELA COM AS REFERÊNCIAS DOS 82 ARTIGOS DO *CORPUS* B – FERREIRA E CORRÊA

Artigo 1 - MASIERO, P. C. *et al*. A Biblioteca Digital de Teses e Dissertações da Universidade de São Paulo. **Ciência da Informação**, Brasília, v. 30, n. 3, p. 34-41, dez. 2001. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652001000300005&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 2 -SILVA, M. C.; SILVA, F. M. A.; AQUINO, M. A. A biblioteca digital paulo freire recuperando o conteúdo freireano para consolidação de políticas de ações afirmativas. **Informação & Sociedade:** Estudos, v. 18, n. 2, 2008. Disponível em: http://hdl.handle.net/20.500.11959/brapci/93225. Acesso em: 16 out. 2021.
Artigo 3 - CUNHA, M. B. da. A biblioteca universitária na encruzilhada. **DataGramaZero: Revista de Ciência da Informação**, Rio de Janeiro, v. 11, n.6, dez. 2010. Disponível em: http://www.dgz.org.br/dez10/Art_07.htm. Acesso em: 16 out. 2021.
Artigo 4 - BOERES, S. A. A.; FARIA, A. C. C. A preservação digital na biblioteca central da universidade de brasília. **Ciência da Informação**, v. 41, n. 1, 2012. DOI: 10.18225/ci.inf..v41i1.1363 Acesso em: 16 out. 2021.
Artigo 5 - FREIRE, I. M. A rede de projetos do núcleo temático da seca da ufrn como possibilidade de socialização da informação. **Informação & Sociedade: Estudos**, v. 14 n.2 2004, n. 2, 2004. Disponível em: http://hdl.handle.net/20.500.11959/brapci/92085. Acesso em: 16 out. 2021.
Artigo 6 - ALVARENGA, L. A teoria do conceito revisitada em conexão com ontologias e metadados no contexto das bibliotecas tradicionais e digitais. **DataGramaZero: Revista de Ciência da Informação**, v. 2, n. 6, 2001. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5300. Acesso em: 16 out. 2021.
Artigo 7 - MOREIRO-GONZÁLEZ, J. A.; ALVES, F. M.; GUAMBE, M. F. Abordagem metodológica para o estudo comparativo entre as bibliotecas digitais em moçambique, brasil e paraguai. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v. 18, n. 37, p. 157-174, 2013. DOI: 10.5007/1518-2924.2013v18n37p157 Acesso em: 16 out. 2021.
Artigo 8 – LIMA, I. F.; SOUZA, R. R.; DIAS, G. A. Abordagens para Avaliar Bibliotecas Digitais. *In*: **XII Encontro Nacional de Pesquisa em Ciência da Informação**. BENANCIB, Florianópolis. 2011. Disponível em: http://repositorios.questoesemrede.uff.br/repositorios/bitstream/handle/123456789/2026/Abordagens%20-%20Lima.pdf?sequence=1. Acesso em: 16 out. 2021.
Artigo 9 – REENEN, J. V. Acesso aberto e conexidade: estimulando inovações inesperadas com o uso de arquivos abertos institucionais. **Ciência da Informação**, v. 35, n. 2, ago. 2006. Disponível em: https://www.scielo.br/j/ci/a/vMqJMTFdxmkdyqJCrWMBXDc/?lang=en&format=pdf. Acesso em: 16 out. 2021.
Artigo 10 - EVANGELISTA, R.; OLIVEIRA, V. F. F.; PEREIRA, S. L.; PETINARI, V. S. Acesso digital: o direito à informação na área da saúde versus a propriedade intelectual da informação tecnológica. **Revista Digital de Biblioteconomia & Ciência da Informação**, v. 3, n. 1, p. 41-66, 2005. DOI: 10.20396/rdbci.v2i2.2065 Acesso em: 16 out. 2021.
Artigo 11 - FREIRE, I. M.; CARVALHO, L. M.; CARVALHO, M. M.; ARANHA, T. Q. Ampliando o acesso livre à informação: a digitalização do acervo do núcleo temático da seca. **Informação & Sociedade: Estudos**, v. 18, n. 2, 2008. Disponível em: http://hdl.handle.net/20.500.11959/brapci/92548. Acesso em: 16 out. 2021.
Artigo 12 - PACHECO, R. C. D. S.; KERN, V. M. Arquitetura conceitual e resultados da

integração de sistemas de informação e gestão da ciência e tecnologia. **DataGramaZero,** v. 4, n. 2, 2003. Disponível em: http://hdl.handle.net/20.500.11959/brapci/3877. Acesso em: 16 out. 2021.
Artigo 13 - CAMARGO, L. S. A.; VIDOTTI, S. A. B. G. Arquitetura da informação para biblioteca digital personalizável 10.5007/1518-2924.2006v11nesp1p103. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, n. esp. 1. sem., p. 103-118, 2006. DOI: 10.5007/1518-2924.2006v11nesp1p103. Acesso em: 16 out. 2021.
Artigo 14 - FUJITA, M. S. L. Aspectos evolutivos das bibliotecas universitárias em ambiente digital na perspectiva da rede de bibliotecas da unesp. **Informação & Sociedade: Estudos**, v. 15 n.2 2005, n. 2, 2005. Disponível em: http://hdl.handle.net/20.500.11959/brapci/91372. Acesso em: 16 out. 2021.
Artigo 15 – PEREIRA, F.; LIMA, G. A. B. O. Avaliação de Usabilidade da Biblioteca digital Brasileira de Teses e Dissertações: um estudo de caso. *In*: **XII Encontro Nacional de Pesquisa em Ciência da Informação**. BENANCIB, Brasília, 2011. Disponível em: http://200.20.0.78/repositorios/bitstream/handle/123456789/2020/Avalia%C3%A7%C3%A3o%20-%20Lima.pdf?sequence=1. Acesso em: 16 out. 2021.
Artigo 16 - CUNHA, M. B. Bibliografia sobre o fluxo do documento na biblioteca digital. **DataGramaZero**, v. 10, n. 5, 2009. Disponível em: http://hdl.handle.net/20.500.11959/brapci/6959. Acesso em: 16 out. 2021.
Artigo 17 - CHATAIGNIER, M. C. P.; SILVA, M. P. Biblioteca digital: a experiência do impa. **Ciência da Informação**, v. 30, n. 3, 2001. DOI: 10.18225/ci.inf..v30i3.907. Acesso em: 16 out. 2021.
Artigo 18 - CUNHA, Murilo Bastos da. Biblioteca digital: bibliografia das principais fontes de informação. Ci. Inf., Brasília , v. 39, n. 1, p. 88-107, abr. 2010 . Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652010000100006&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 19 - SOUTHWICK, Sílvia Barcellos. The Brazilian electronic theses and dissertations digital library: providing open access for scholarly information. **Ciência da Informação**, Brasília, v. 35, n. 2, p. 103-110, aug. 2006. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652006000200011&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 20 - Santos Filho, J. M. dos, & Kaimen, M. J. G.-. (2009). Biblioteca digital como recurso informacional no ensino a distância (EaD): uma análise das instituições de ensino superior (IESs) credenciadas para programas de EaD na região Sul do país. **Informação & Sociedade: Estudos**, v. 19, n. 3. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/2390. Acesso em: 16 out. 2021.
Artigo 21 - SILVA, Angela Maria; SILVA, Ilmério Reis; ARANTES, Luiz Humberto Martins. Biblioteca digital de peças teatrais. **Ciência da Informação**, Brasília, v. 33, n. 2, p. 187-196, aug. 2004. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652004000200020&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 22 - BOTTARI, Christina Thereza Rachel; DA SILVA, Neusa Cardim. Biblioteca digital de teses e dissertações da UERJ: desafios e oportunidades. **Informação & Informação**, v. 16, n. 1, p. 88-101, 2011. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/7091. Acesso em: 16 out. 2021.
Artigo 23 - FERNÁNDEZ, Juan J.; MATIAS, Marcio. Biblioteca digital de teses e dissertações do IBICT: uma análise sob a ótica da arquitetura da informação. **Revista ACB: Biblioteconomia em Santa Catarina**, v. 22, n. 2, p. 285-299, 2017. Disponível em: https://revista.acbsc.org.br/racb/article/view/1346. Acesso em: 16 out. 2021.

Artigo 24 - MORAIS, P. S.; PINHEIRO, E. G. Biblioteca digital paulo freite à luz dos direitos autorais: um sonho a mais não faz mal. **Biblionline**, v. 1, n. 2, 2005. Disponível em: http://hdl.handle.net/20.500.11959/brapci/16693. Acesso em: 16 out. 2021.
Artigo 25 - VITORINO, Elizete Vieira; ISAMI, Brenda Dayana Gonzalez. Biblioteca digital sobre Educação a Distância (EaD): favorecendo o acesso ao acervo do Núcleo de Estudos e Pesquisas em Competência Informacional (GPCIN). **Revista ACB,** Santa Catarina, v. 18, n. 1, p. 531-552, 2013. Disponível em: file:///C:/Users/marci/Downloads/861-4109-1-PB.pdf. Acesso em: 16 out. 2021.
Artigo 26 - SOARES, A. A. O.; FARIA, F. M. D. S.; MENDES, G. O.; ARAúJO, G. L. A. L.; DIAS, M. F.; SILVA, R. I. V.; GONçALVES, R. C. Biblioteca digital universal. **Múltiplos Olhares em Ciência da Informação**, v. 4, n. 2, 2014. Disponível em: http://hdl.handle.net/20.500.11959/brapci/64292. Acesso em: 16 out. 2021.
Artigo 27 - SALVIATI, Maria Elisabeth; DUARTE, DH de O. Biblioteca Eletrônica da Embrapa Cerrados: estudo de usuário. **Embrapa Cerrados-Artigo em periódico indexado (ALICE)**, 2015. Disponível em: https://www.alice.cnptia.embrapa.br/handle/doc/1078487. Acesso em: 16 out. 2021.
Artigo 28 - ZAFALON, Z. R. Biblioteca em tempo real: o acesso em foco: proposta crítica do modelo de organização da informação na contemporaneidade. **Revista Digital de Biblioteconomia & Ciência da Informação**, v. 6, n. 2, p. 61-83, 2008. DOI: 10.20396/rdbci.v6i1.1998 Acesso em: 16 out. 2021.
Artigo 29 - GOMES, S. L. R. Biblioteca virtual: um novo território para a pesquisa científica no brasil. **DataGramaZero**, v. 5, n. 6, 2004. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5715. Acesso em: 16 out. 2021.
Artigo 30 - ROSETTO, Marcia. Bibliotecas Digitais–Cenário e Perspectivas. **Revista Brasileira de Biblioteconomia e Documentação**, v. 4, n. 1, p. 101-130, 2008. Disponível em: https://rbbd.febab.org.br/rbbd/article/view/101/92. Acesso em: 16 out. 2021.
Artigo 31 - ALENCAR, A. F. Bibliotecas digitais: uma nova aproximação. **Informação & Sociedade: Estudos**, v. 14 n.1 2004, n. 1, 2004. Disponível em: http://hdl.handle.net/20.500.11959/brapci/90992. Acesso em: 16 out. 2021.
Artigo 32 - SCHMIDT, Luciana; OHIRA, Maria Lourdes Blatt. Bibliotecas virtuais e digitais: análise das comunicações em eventos científicos (1995/2000) Virtual and digital libraries: analysis communications in scientific events (1995-2000) p. 73-97. **Revista ACB**, v. 7, n. 1, p. 73-97, 2002. Disponível em: https://revista.acbsc.org.br/racb/article/view/377/455. Acesso em: 16 out. 2021.
Artigo 33 - OHIRA, Maria Lourdes Blatt; PRADO, Noêmia Schoffen. Bibliotecas virtuais e digitais: análise de artigos de periódicos brasileiros (1995/2000). **Ciência da Informação**, Brasília, v. 31, n. 1, p. 61-74, jan. 2002. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652002000100007&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 34 - PRYSTHON, Cecília; SCHMIDT, Susana. Experiência do Leaal/UFPE na produção e transferência de tecnologia. **Ciência da informação**, v. 31, p. 84-90, 2002. Disponível em: http://revista.ibict.br/ciinf/article/view/980/0. Acesso em: 16 out. 2021.
Artigo 35 - DE SALES, Rodrigo. Colóquio Internacional Bibliotecas Digitais–Brasil–França–Alemanha: relato de experiência Colloquium of International Digital Libraries: Brazil-France–Germany: report p. 353-362. **Revista ACB**, v. 11, n. 2, p. 353-362, 2006. Disponível: https://revista.acbsc.org.br/racb/article/view/457. Acesso em: 16 out. 2021.
Artigo 36 - BOERES, S. A. A.; CUNHA, M. B. Competências básicas para os gestores de preservação digital. **Ciência da Informação**, v. 41, n. 1, 2012. DOI: 10.18225/ci.inf..v41i1.1356 Acesso em: 16 out. 2021.
Artigo 37 - CUNHA, Murilo Bastos da. Das bibliotecas convencionais às digitais: diferenças

e convergências. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 13, n. 1, p. 2-17, abr. 2008. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362008000100002&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 38 - GREENHALGH, Raphael Diego. Digitalização de obras raras: algumas considerações. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 16, n. 3, p. 159-167, set. 2011. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362011000300010&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 39 - FERREIRA, S. M. S. P.; GADELHA, Z.; GAMBA, C. M. Digitalização e preservação digital: a experiência do sistema integrado de bibliotecas da universidade de são paulo (sibiusp). **Ciência da Informação**, v. 41, n. 1, 2012. DOI: 10.18225/ci.inf..v41i1.1360. Acesso em: 16 out. 2021.
Artigo 40 - SANTOS, Gildenir Carolino; PASSOS, Rosemary. Estratégias para a estruturação de um website visando o desenvolvimento de Bibliotecas Digitais. **ETD: Educaçao Temática Digital**, v. 6, n. 1, p. 59-67, 2004. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/etd/article/view/1002. Acesso em: 16 out. 2021.
Artigo 41 - MARCIAL, E. C.; PINHEIRO, C. C. M.; NASCIMENTO, M. E. M.; MENEZES, J. T. M. Estudo de viabilidade de rede de bibliotecas em nuvem da administração pública federal. **Revista Ibero-Americana de Ciência da Informação**, v. 9, n. 1, p. 143-161, 2016. DOI: 10.26512/rici.v9.n1.2016.2209 Acesso em: 16 out. 2021.
Artigo 42 - FOX, Edward A.; YANG, Seungwon; KIM, Seonho. ETDs, NDLTD, and open access: a 5S perspective. **Ciência da Informação**, Brasília , v. 35, n. 2, p. 75-90, Aug. 2006. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652006000200009&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 43 - DA ROCHA MIRANDA, Roberto Campos; TEIXEIRA, Sandra C.; FILIZOLA, Adriana R. Gestão do conhecimento aplicada a bibliotecas digitais: estudo de caso no Ministério da Saúde e na Infraero. **Brazilian Journal of Information Science**, v. 10, n. 1, p. 49-55, 2016. Disponível em: https://dialnet.unirioja.es/servlet/articulo?codigo=5376139. Acesso em: 16 out. 2021.
Artigo 44 - GONZALEZ, Marco; POHLMANN FILHO, Omer; BORGES, Karen Selbach. Informação digital no ensino presencial e no ensino a distância. **Ciência da Informação**, Brasília, v. 30, n. 2, p. 101-111, aug. 2001. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652001000200012&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 45 - MARCONDES, Carlos Henrique; SAYAO, Luís Fernando. Integração e interoperabilidade no acesso a recursos informacionais eletrônicos em C&T: a proposta da Biblioteca Digital Brasileira. **Ciência da Informação**, Brasília, v. 30, n. 3, p. 24-33, dez. 2001. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652001000300004&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 46 - LIMA, I. F.; SOUZA, R. R.; DIAS, G. A. Interatividade e usabilidade nas bibliotecas digitais no processo ensino-aprendizagem. **DataGramaZero**, v. 13, n. 3, 2012. Disponível em: http://hdl.handle.net/20.500.11959/brapci/7843. Acesso em: 16 out. 2021.
Artigo 47 - DE ALMEIDA RODRIGUES, Nelson. Introdução ao METS–Preservação e Intercâmbio de Objetos Digitais. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 13, n. 26, p. 172-187, 2008.
Artigo 48 - SANTOS, Gildenir Carolino. Mapeamento dos suportes de auxílio ao ensino tradicional: uma contextualização da biblioteca, do livro, do computador, da internet e da tecnologia na educação. **ETD-Educação Temática Digital**, v. 4, n. 2, p. 48-62, 2003.

Disponível em: https://www.ssoar.info/ssoar/handle/document/10423. Acesso em: 16 out. 2021.
Artigo 49 - PONTES, Flávio Vieira; LIMA, Gercina Ângela Borém de Oliveira. Modelos Conceituais para Bibliotecas Digitais. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**, v. 8, n. 2, 2014. Disponível em: https://www.ssoar.info/ssoar/handle/document/10423. Acesso em: 16 out. 2021.
Artigo 50 - CASTRO, Jetur Lima de *et al*. Preservação digital em coleções bibliográficas da biodiversidade: o caso da Biodiversity Heritage Library no Museu Paraense Emílio Goeldi (MPEG). **Revista Digital de Biblioteconomia e Ciência da Informação**, v. 1, n. 33, 2016.
Artigo 51 - DE CARVALHO, Maria Carmen Romcy; DE TOLEDO DUBOIS, Maria Célia; COVÕES, Thiago Ferreira. O acesso aberto à produção científica das universidades católicas: o caso da CVA-RICESU. **Encontros Bibli,** n. Especial 1, p. 95-103, 2007.
Artigo 52 - COSTA, Maira Murrieta; CUNHA, Murilo Bastos da. O bibliotecário no tratamento de dados oriundos da e-science: considerações iniciais**. Perspectivas em Ciência da Informação**, v. 19, n. 3, p. 189-206, 2014.
Artigo 53 - LUCAS, Clarinda Rodrigues. O conceito de biblioteca nas bibliotecas digitais. **Informação & Sociedade**, v. 14, n. 2, 2004. Disponível em: https://www.brapci.inf.br/_repositorio/2010/11/pdf_602d3acd9a_0013008.pdf. Acesso em: 16 out. 2021.
Artigo 54 - CASTRO, B. O.; CUNHA, M. B. O ensino da biblioteca digital nos currículos de graduação em biblioteconomia. **Revista Digital de Biblioteconomia & Ciência da Informação**, v. 11, n. 2, p. 197-221, 2013. DOI: 10.20396/rdbci.v11i2.1645 Acesso em: 19 out. 2021.
Artigo 55 - ASSUNCAO, Renato Vieira da; REIS, Cley Arthur Miranda. O Futuro das bibliotecas pós-Google Books. **Pesquisa Brasileira em Ciência da Informação e Biblioteconomia**; v. 8, n. 1, 2013.
Artigo 56 - NARDINO, Anelise Tolotti Dias; CAREGNATO, Sônia Elisa. O futuro dos livros do passado: a biblioteca digital contribuindo na preservação e acesso às obras raras. **Em Questão**, v. 11, n. 2, p. 381-407, 2005.
Artigo 57 - ODDONE, N. E.; MEIRELLES, R. F. O portal de periódicos da capes e os indicadores de desempenho da informação eletrônica. **DataGramaZero**, v. 7, n. 3, 2006. Disponível em: http://hdl.handle.net/20.500.11959/brapci/5917. Acesso em: 19 out. 2021.
Artigo 58 - ALVES, Ana Paula Meneses; VIDOTTI, Silvana Aparecida Borsetti Gregório. O serviço de referência e informação digital. **Biblionline**, v. 2, n. 2, p. p1-10, 2006.
Artigo 59 - MARTINS, Robson Dias. Perspectivas para uma biblioteca no futuro: utopia ou realidade. **Informação & Sociedade**, v. 12, n. 1, 2002.
Artigo 60 - PAES, Denyse Maria Borges; TABOSA, Hamilton Rodrigues. **Biblioteca Digital de Teses e Dissertações**: reflexões sobre representação da informação com vistas à recuperação da informação. Revista ACB, Santa Catarina, v. 20, n. 2, 2015. Disponível em: https://revista.acbsc.org.br/racb/article/view/1007. Acesso em: 16 out. 2021.
Artigo 61 - FERREIRA, Sueli Mara Soares Pinto *et al.* **Da política institucional de informação da Universidade de São Paulo ao acesso aberto à produção científica do Cruesp**. Reciis, v. 8, n. 2, 2014. . Disponível em: https://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/632. Acesso em: 16 out. 2021.
Artigo 62 - DE SOUSA SANTOS, Rayane Soares; DE GÓES BRENNAND, Edna Gusmão. DOCUMENTOS DIGITAIS E DIREITOS AUTORAIS: reflexões na Biblioteca Digital Paulo Freire. **Ponto de Acesso**, v. 9, n. 2, p. 65-83, 2015. Disponível em: https://brapci.inf.br/index.php/res/download/98727. Acesso em: 16 out. 2021.
Artigo 63 - PAIS, Clarisse *et al*. Serão as políticas institucionais mandatórias, assim tão mandatórias? Qual o grau de cumprimento? O caso da Biblioteca Digital do IPB. **Ponto de**

Acesso, v. 9, p. 3-17, 2015.
Artigo 64 - BRUMATTI, Josimara Dias. A contribuição da Biblioteca Digital de Teses e Dissertações na disseminação do conhecimento nas áreas de Humanas e Sociais. **Revista Brasileira de Biblioteconomia e Documentação**, v. 11, n. 1, p. 66-77, 2015.
Artigo 65 - MARTINS, Dalton Lopes; FERREIRA, Sueli Mara Soares Pinto. Análise da Dinâmica de Evolução das revistas científicas e Bibliotecas Digitais de Teses e Dissertações em acesso livre na área das Ciências da Comunicação: o caso do repositório univerciencia. **Encontros Bibli**, v. 17, n. 2, p. 136-158, 2012.
Artigo 66 - SCHWEITZER, Fernanda; RODRIGUES, Rosângela Schwarz. Teses e dissertações em tecnologias de informação e comunicação integradas com a educação: uma análise da BDTD do IBICT Theses and Dissertations in information and communication with integrated education: an analysis of DLTD IBICT. **Revista ACB**, v. 15, n. 2, p. 90-111, 2010.
Artigo 67 - PAVANI, Ana. A model of multilingual digital library. **Ciência da informação**, v. 30, n. 3, p. 73-81, 2001.
Artigo 68 - SILVA, Marcel Santos; VIDOTTI, Silvana Aparecida Borsetti Gregório. Biblioteca digital geográfica distribuída: uma arquitetura para desenvolvimento. **Informação & Informação**, v. 12, n. 2, p. 150-167, 2007.
Artigo 69 - MIGUEL-ÁNGEL, Marzal-García-Quismondo; AURORA, Cuevas-Cerveró. Biblioteca escolar para la sociedad del conocimiento en España. **Ciência da Informação**, v. 36, n. 1, p. 54-68, 2007.
Artigo 70 - ARELLANO, Miguel Angel. Preservação de documentos digitais**. Ciência da Informação**, v. 33, n. 2, p. 15-27, 2004.
Artigo 71 - MIGUEIS, Ana; FIOLHAIS, Carlos. Recursos digitais em livre acesso na Universidade de Coimbra: Estudo Geral e Alma Mater. **Revista Eletrônica de Comunicação, Informação e Inovação em Saúd**, v. 8, n. 2, p. 231-242, 2014.
Artigo 72 - ALVARENGA, Lídia. Representação do conhecimento na perspectiva da ciência da informação em tempo e espaço digitais. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 8, n. 15, p. 18-40, 2003.
Artigo 73 - MÁRDERO ARELLANO, Miguel Ángel. Serviços de referência virtual. **Ciência da informação**, v. 30, n. 2, p. 7-15, 2001.
Artigo 74 - DOS SANTOS MACULAN, Benildes Coura Moreira; DE OLIVEIRA LIMA, Gercina Angela Borém; PENIDO, Patrícia. Taxonomia facetada como interface para facilitar o acesso à informação em bibilotecas digitais Faceted taxonomy as interface to information's access facility in digital library. **Revista ACB**, v. 16, n. 1, p. 234-249, 2011.
Artigo 75 - GONCALVES, Marcos André; FOX, Edward A.. Technology and research in a global Networked University Digital Library (NUDL)Technology and research in a global Networked University Digital Library (NUDL). **Ciência da Informação**, Brasília , v. 30, n. 3, p. 13-23, dez. 2001.
Artigo 76 - ROSALES, N. Fabiola et al . Tesis electrónicas de la Universidad de Los Andes: adaptación y uso de la Plataforma Tede. **Ciência da Informação,** Brasília , v. 35, n. 2, p. 111-116, aug. 2006 . Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652006000200012&lng=en&nrm=iso. Acesso em: 16 out. 2021.
Artigo 77- AUTRAN, Marynice de Medeiros Matos et al. A transferência do conhecimento para o setor produtivo: experiência de uma parceria. **Biblionline**, João Pessoa, v. 4, n. 1-2, 2008.
Artigo 78- PACHECO, Roberto Carlos dos Santos; KERN, Vinícius Medina. Transparência e gestão do conhecimento por meio de um banco de teses e dissertações: a experiência do PPGEP/UFSC. **Ciência da informação**, v. 30, n. 3, p. 64-72, 2001.
Artigo 79- BOHMERWALD, Paula. Uma proposta metodológica para avaliação de

bibliotecas digitais: usabilidade e comportamento de busca por informação na Biblioteca Digital da PUC-Minas. **Ciência da Informação**, v. 34, n. 1, p. 95-103, 2005.
Artigo 80- CARRARE, Ana Paula et al. Uma proposta para gerenciamento e preservação de imagens em medicina na EPM/Unifesp. **Ciência da Informação**, v. 35, n. 3, p. 201-208, 2006.