Artigo Original

Encontros Bibli

SISTEMATIZAÇÃO DA OBTENÇÃO DE INDICADORES TEMÁTICOS DE INFORMAÇÃO CIENTÍFICA

Systematization of obtaining thematic indicators of scientific information

Márcio Henrique Wanderley Ferreira
Universidade Federal de Pernambuco, Programa de Pós-Graduação em Ciência da Informação, Recife, Brasil
marcio.wferreira@gmail.com
https://orcid.org/0000-0002-2552-325X

Renato Fernandes Correa
Universidade Federal de Pernambuco, Programa de Pós-Graduação em Ciência da Informação, Recife, Brasil
renato.correa@ufpe.br
https://orcid.org/0000-0002-9880-8678

A lista completa com informações dos autores está no final do artigo

RESUMO

Objetivo: No contexto de desenvolvimento de estudos métricos da informação, este trabalho propõe e aplica um método para obtenção de indicadores temáticos sobre descritores representativos de temas, assuntos ou palavras-chave abordados em registros bibliográficos da área de Ciência da Informação.
Método: Realizou-se uma pesquisa metodológica de natureza aplicada, utilizando procedimentos técnicos da indexação automática e dos estudos métricos da informação. Inicialmente, delimitou-se um *corpus* contemplando um conjunto de registros bibliográficos referentes a 60 artigos de periódicos brasileiros. Posteriormente, aplicou-se o *software* Maui como sistema de indexação automática na categorização das palavras-chave dos registros bibliográficos em conceitos de um tesauro de especialidade, contemplando descritores no idioma do texto dos metadados. Em seguida, aplicou-se o *software* Iramuteq, para gerar os indicadores temáticos a partir dos descritores obtidos pela indexação automática. Validou-se, por fim, o método proposto com base na análise dos resultados obtidos para o *corpus*.
Resultado: São descritos os fluxogramas de validação da indexação automática e de validação do estudo métrico, visando identificar e descrever os processos do método proposto como etapas do processamento do *corpus*. Outrossim, apresentam-se as métricas de qualidade na indexação automática, bem como análises estatísticas de frequência e coocorrência de palavras, e de frequência de termos, onde apontam-se os principais indicadores temáticos do *corpus*.
Conclusões: Conclui-se que os indicadores temáticos obtidos por meio da aplicação do método proposto representam os principais temas identificados no *corpus*, e que o método pode ser aplicado na obtenção de indicadores temáticos de outros conjuntos de registros bibliográficos.
PALAVRAS-CHAVE: Estudos métricos da informação. Indexação automática. Indicadores temáticos. Bibliometria. Trajetória metodológica.

ABSTRACT

Objective: In the context of the development of information metric studies, this article proposes and applies a method for obtaining thematic indicators on descriptors representative of themes, subjects or keywords addressed in bibliographic records in the Information Science area.
Method: A methodological research of applied nature was carried out, using technical procedures of automatic indexing and information metric studies. Initially, a corpus was delimited contemplating a set of bibliographic records referring to 60 articles of Brazilian journals. Later, the software Maui was applied as an automatic indexing system for categorizing the keywords of the bibliographic records into concepts of a specialty thesaurus, contemplating descriptors in the language of the metadata text. Next, the Iramuteq software was applied to generate the thematic indicators from the descriptors obtained by the automatic indexing. Finally, the proposed method was validated based on the analysis of the results obtained for the corpus.
Result: The flowcharts for validation of automatic indexing and for validation of the information metric study are described, aiming to explain the processes of the proposed method as steps in processing the corpus. In addition, quality metrics for automatic indexing are presented, as well as statistical analyses of word frequency and co-occurrence, and term frequency, where the corpus' main thematic indicators are pointed out.
Conclusions: We conclude that the thematic indicators obtained by applying the proposed method represent the main themes identified in the corpus, and that the method can be applied to obtain thematic indicators for other sets of bibliographic records.
KEYWORDS: Information metric studies. Automatic indexing. Thematic indicators. Bibliometrics. Methodological pathway.

1 INTRODUÇÃO

Segundo Oliveira e Grácio (2011) os estudos métricos da informação surgem da união de disciplinas como Bibliometria, Cientometria, Webometria, Infometria e Altmetria, sendo esta última a mais recentemente congregada. Nesse sentido, é possível afirmar que o desenvolvimento de estudos métricos da informação encontra instrumentos apropriados na Ciência da Informação (LE COADIC, 2004).

Os estudos métricos da informação possibilitam a descrição e análise da produção científica de diversas áreas, campos e domínios do conhecimento. Tais estudos fazem uso de instrumentos capazes de identificar realidades antes desconhecidas, provocando reflexões sobre os comportamentos científicos individuais e coletivos, visíveis apenas pela aplicação das métricas informacionais, indicando conjunturas e cenários por meio da análise de indicadores científicos previamente determinados.

Dentro do escopo de desenvolvimento de estudos métricos da informação, a formulação de indicadores temáticos constitui um caminho válido para contribuir no processo de entendimento acerca de um determinado *corpus* bibliográfico*,* uma vez que possuem a capacidade de representar o conhecimento sobre os temas tratados nos documentos, por meio de visualizações gráficas.

No escopo deste trabalho, podemos definir o indicador temático como sendo um indicador de ligação da atividade científica que retrata associações temáticas que descrevem a força da relação de frequência de ocorrência ou coocorrência entre temas associados às referências bibliográficas, levando a um maior entendimento sobre as temáticas principais ou a base conceitual abordada nas publicações. Dentro do arcabouço teórico da Bibliometria, a Lei de Zipf tem importante papel na formulação de tais indicadores, pois, aplicando-a, é possível identificar a frequência de ocorrência de uma dada palavra, e determinar associações temáticas (BEIRA *et al*., 2020).

Contudo definir quais termos representam, de forma mais precisa, determinado conjunto de documentos, não é uma tarefa simples, exigindo do pesquisador conhecimento e prática de indexação (SANTOS, 2015a). Além disso, o processo de indexar um conjunto de documentos pode exigir um longo tempo de atividade intelectual, mesmo antes da construção de indicadores temáticos. Em contrapartida, o uso das palavras-chave do autor como temas, pode gerar indicadores temáticos pouco representativos, devido a dispersão terminológica característica do uso de termos em linguagem natural e a subjetividade envolvida na atribuição de tais termos.

Alternativamente, a indexação pode ser realizada por meio da indexação automática. A indexação automática é definida por Corrêa e Lapa (2013, p. 258) 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.

A indexação automática visa tornar o processo de indexação mais rápido e menos custoso, agilizando o processo de atribuição de termos relevantes aos documentos digitais, por meio do processamento computacional dos termos presentes no conteúdo textual de títulos, resumos e textos completos. Quando um tesauro eletrônico é utilizado, se está realizando a indexação automática por atribuição, caso contrário, se está realizando a indexação automática por extração. Por instância, alguns trabalhos recentes têm investigado a indexação automática por atribuição de artigos de periódicos da área de Ciência da Informação escritos no idioma português do Brasil (BANDIM; CORREA, 2018, 2019; SILVA; CORREA; GIL-LEIVA, 2020). Adicionalmente, para além do escopo do presente artigo, recomenda-se o artigo de Gil-Leiva, Ortuño e Corrêa (2022) para a leitura de uma revisão bibliográfica internacional e aprofundada sobre a indexação automática de artigos científicos.

Diante da problemática da formulação de indicadores temáticos e das potencialidades apresentadas pela combinação das técnicas de indexação automática e de estudos métricos da informação para o alcance de soluções, este artigo parte do seguinte problema de pesquisa: é possível estabelecer um método para sistematizar a formulação de indicadores temáticos de informação científica? A fim de responder a esse questionamento, o objetivo geral deste trabalho é propor um método para a formulação de indicadores temáticos de informação científica da área de Ciência da Informação no Brasil, pautando-se em trajetória metodológica a ser realizada empiricamente pelo pesquisador para alcance de tais indicadores.

A justificativa deste estudo baseia-se na concepção da proposição de um método, que aplica técnicas de indexação automática por atribuição e de estudos métricos da informação, com foco na geração de indicadores temáticos de um conjunto de registros bibliográficos da área da Ciência da Informação no Brasil. Almeja-se empregar o método proposto na classificação do conhecimento científico e na produção de estudos metacientíficos, a fim de retratar e analisar o conhecimento registrado em textos científicos.

2 ESTUDOS MÉTRICOS DA INFORMAÇÃO

Noronha e Maricato (2008) analisam a evolução dos estudos métricos da informação e categorizam as técnicas que podem ser empregadas nos mesmos, de acordo com o objeto de estudo e finalidade. Dentre as técnicas, a mais antiga é a Bibliometria, sendo também reconhecida como disciplina por outros autores.

Alvarado (1984) afirma que a Bibliometria se originou no início do século XX, a partir de um processo de medição do conhecimento por meio da realização de estudos matemáticos e da utilização de técnicas experimentais baseadas em métodos quantitativos. A Bibliometria contempla análise da comunidade científica e sua estrutura, revelando aspectos das redes de pesquisadores e suas respectivas motivações. Estuda, ainda, a documentação, mediante contagem de autores e periódicos, podendo identificar o núcleo e a periferia da produção acadêmica. Adicionalmente, examina indicadores de produtividade e de qualidade científica e tecnológica (OKUBO, 1997), visando principalmente, a avaliação de instituições e da produtividade de autores, bem como o ranqueamento de revistas (ARAÚJO, 2018).

Nesse cerne, teorias e leis foram criadas para compor a análise dos fenômenos específicos do campo em questão. Dentre as principais leis, destacam-se: a lei de Bradford sobre a produtividade de periódicos; a lei de Lotka sobre a produtividade de autores; e a lei de Zipf sobre a frequência de ocorrência de palavras.

A lei de Zipf aponta para a existência de uma economia no uso de palavras pelos pesquisadores, bem como para o fato de que palavras significativas mais frequentes indicam assuntos do documento (ALVARADO, 2007). A referida lei fornece uma das explicações estatísticas para que os sistemas de indexação automática, por exemplo, escolham termos mais frequentes, após a remoção das palavras vazias de significado *(stopwords*) e normalização de palavras, como representativos do conteúdo semântico do documento. Por isso, a lei de Zipf é adotada como fundamento para o desenvolvimento do método proposto neste artigo.

De acordo com Kobashi e Santos (2006, p. 32) 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); e indicadores de ligação (coocorrência de autoria, citações e palavras). Nesta pesquisa, utiliza-se a noção de indicadores temáticos com sendo

indicadores de ligação baseados na frequência de ocorrência e coocorrência de palavras, que explicitam os temas abordados em um conjunto de registros bibliográficos.

2.1 Indicadores temáticos

A cartografia bibliométrica (do inglês *bibliometric cartography*) foi aplicada por Noyons e Van Raan (1994) como método analítico para o estudo dos aspectos importantes relacionados à ciência e à tecnologia no campo da optomecatrônica. Nessa pesquisa, os autores desenvolveram mapas da ciência, baseando-se na coocorrência de palavras-chave de publicações e patentes.

Segundo os supracitados pesquisadores, as principais vantagens da utilização da representação por cartografia bibliométrica são: a visualização oferece um panorama em menos tempo, pois é mais facilmente assimilada e memorizada; e tal representação sintetiza e reduz a informação, permitindo a filtragem de características significativas. Noyons e Van Raan (1994) afirmam, ainda, que a utilização de uma abordagem por cartografia facilita a visualização por subcampos semelhantes, com agrupamentos de copalavras. Dessa maneira, a construção desses mapas permite uma melhor definição das ligações entre os termos analisados.

De acordo com Ding, Chowdhury e Foo (1999) a cartografia bibliométrica é um método de visualização de cocitação em um conjunto de documentos. Na análise de cocitação, os dados são projetados e reduzidos em uma representação visual específica, com a manutenção das informações essenciais contidas nos dados. Nesse método, é utilizada a análise de escalonamento multidimensional (do inglês *Multidimensional Scaling* – MDS) para criar a visualização em formato de mapa, a partir de matriz de cocitação. Desse modo, essa técnica, originalmente desenvolvida para representar visualmente a cocitação de autores, foi gradualmente sendo utilizada para outros tipos de análise de cocitação, como a de copalavras (do inglês *co-word analysis*).

Segundo Kobashi e Santos (2006) indicadores de ligação, como a coocorrência de palavras, são indicadores temáticos levantados pelo método de cartografia temática (do inglês *thematic cartography*), que representam padrões de comportamento nos campos científicos, identificando o conhecimento disseminado ao longo do tempo e suas possíveis relações e dinâmicas. Tendo como base o trabalho dos referidos autores, pode-se definir os mapas ou gráficos de temas como visualizações gráficas que permitem ao ser humano identificar, de forma mais global e compreensível, os assuntos tratados em um conjunto de registros bibliográficos, tendo respaldo em estudos sobre a percepção humana.

Posteriormente, Kobashi e Santos (2008) realizaram a cartografia temática de teses e dissertações da área de pesquisas nucleares. Dentre os trabalhos que mais recentemente realizaram a cartografia temática em publicações escritas no idioma português do Brasil, destacam-se: a utilização por Kobashi, Díaz e Santana (2014) da cartografia temática para gerar indicadores temáticos sobre o tema *organização da informação*; e a realização por Pinto *et al.* (2017) da cartografia temática da produção intelectual da Embrapa, destinada à agricultura familiar e registrada nas publicações técnico-científicas editadas pela empresa.

Outro método relacionado ao levantamento de indicadores temáticos é o mapeamento bibliométrico (do inglês *bibliometric mapping*), mais especificamente na elaboração de mapas de termos (VAN ECK *et al.*, 2010a), que facilitam a visualização dos principais assuntos de um domínio analisado. Segundo Van Eck e Waltman (2010), o mapeamento bibliométrico é um importante tópico de pesquisa no campo da bibliometria (BORNER; CHEN; BOYACK, 2003), apresentando-se como uma ferramenta poderosa para estudar e analisar a dinâmica dos campos científicos e compreender melhor uma determinada área de pesquisa.

Nessa seara, Van Eck e Waltman (2010) apresentam o *software VOSViewer*[1], que é voltado para a representação gráfica de mapas bibliométricos. O *software* em questão permite construir mapas de palavras-chave com base em dados de coocorrência, gerando mapas bibliométricos por escalonamento multidimensional (MDS) e pela técnica de visualização de semelhanças (VOS). Em Van Eck *et al.* (2010b) é feita uma comparação entre as duas técnicas, que, embora gerem mapas diferentes, apresentam o mesmo objetivo, que consiste em representar graficamente a proximidade e a distância entre os itens descritivos de um documento. Nessa linha de pesquisa, visando a contribuir para a visualização do desenvolvimento científico da Arquivologia no Brasil, recentemente, Rodrigues, Azevedo e Batalha (2021) fizeram uso do *VOSViewer* para construir mapas bibliométricos de palavras-chave de trabalhos em anais de oito edições do Congresso Nacional de Arquivologia.

Há também um estudo desenvolvido por Santos (2015b) em que a autora realizou a bibliometria temática de artigos indexados na Base de Dados em Ciência da Informação (BRAPCI) sobre organização e representação do conhecimento, publicados entre 1996 e 2013. A pesquisadora aplicou o *software* NVivo[2] para gerar as análises de agrupamento e

[1] Disponível em: https://www.vosviewer.com/
[2] Disponível em: https://lumivero.com/products/nvivo/

identificar as similaridades semânticas entre os termos e as correlações temáticas na área de organização e representação do conhecimento.

Em síntese, independentemente da denominação do método bibliométrico de geração de indicadores temáticos e suas respectivas visualizações, os trabalhos descritos apontam as vantagens do levantamento de indicadores temáticos e da representação cartográfica de temáticas na visualização da informação e análise bibliométrica, representando a dinâmica em torno dos assuntos mais discutidos e suas relações.

Além disso, no melhor conhecimento dos autores, corroborado por busca realizada na base Web of Science no ano de 2022, não se sabe da existência de método semelhante ao proposto neste artigo, pautado na combinação de técnicas da indexação automática por atribuição e da análise bibliométrica de frequência de ocorrência e coocorrência, visando a sistematização da formulação de indicadores temáticos de informação científica.

Foi utilizada a seguinte expressão de busca por tópicos na base: ("automatic indexing" OR "keyword extraction" OR "keyphrase extraction") AND ("thematic cartography" OR “bibliometric cartography" OR "thematic bibliometry" OR "bibliometric mapping" OR “bibliometric” OR “bibliometry”). Foram recuperados nove trabalhos, dos quais três apresentam como semelhança a aplicação de sistema de indexação automática na realização de estudo métrico, a saber: em (BHUYAN; SANGURI; SHARMA, 2021) foi aplicado o algoritmo RAKE em resumos, e depois foi aplicada a análise de matriz de coocorrência para agrupamento das palavras-chave extraídas; em (DOLOREUX *et. al*., 2019) foi aplicado o RAKE no texto completo de artigos para posteriormente identificar associações entre as palavras-chave extraídas; e em (OH; LEE, 2014) foi aplicado o sistema KEA[3] em resumos para extrair palavras-chaves do autor, para depois realizar a análise de coocorrência dos termos extraídos. Diferentemente do método proposto, os três trabalhos realizaram a indexação automática por extração (e não por atribuição), e as palavras-chave extraídas foram consideradas sem passar por validação, na análise de coocorrência para fins de agrupamento de temas.

3 METODOLOGIA

Baseando-se na categorização proposta por Vergara (2007), no presente artigo, realizou-se uma pesquisa metodológica, que compreende estudo que se refere ao desenvolvimento de instrumentos de captação ou manipulação da realidade, sendo

[3] Disponível em http://community.nzdl.org/kea/index.html

associado, portanto, a caminhos, formas, maneiras e procedimentos para atingir determinado fim. Nesse contexto, este trabalho propõe um método para a formulação de indicadores temáticos de informação científica a partir de um conjunto de registros bibliográficos da área de Ciência da Informação, com aportes da indexação automática e dos estudos métricos da informação.

Quanto aos meios, constitui-se em pesquisa bibliográfica, para apresentar e sistematizar os fundamentos e as experiências presentes na literatura, valendo-se de material já publicado e revisado por pares. Constitui-se também em pesquisa empírica, por basear-se em evidências e resultados obtidos por um conjunto de procedimentos qualiquantitativos, envolvendo aplicação de recursos algorítmicos para processamento de dados e produção de estatísticas. Além disso, a pesquisa tem natureza aplicada, por buscar resolver o problema da sistematização da formulação de indicadores temáticos.

Os principais aportes técnicos e teóricos selecionados estão associados à organização da informação, que discute os princípios da indexação automática, e aos estudos métricos da informação, com ênfase na produção de indicadores temáticos sobre a produção de conhecimento. Adicionalmente, o *Business Process Management* (BPM) foi utilizado para representar fluxos e processos do método proposto.

Inicialmente, a pesquisa adotou na construção do método proposto de duas etapas: a aplicação de um sistema de indexação automática que precede a realização de estudo métrico com a aplicação de um software de análise bibliométrica. O sistema de indexação automática foi aplicado com o intuito de realizar a representação temática de um conjunto de registros bibliográficos. Em seguida, na procura por um enfoque representativo, realizou-se um estudo métrico da informação pautado na análise de gráficos de frequência e de similitude, fazendo-se uso de técnicas estatísticas para a análise bibliométrica dos termos de indexação obtidos via indexação automática, com o intuito de realizar um levantamento de indicadores temáticos de um conjunto de registros bibliográficos.

Posteriormente, para a validação do método proposto de forma empírica, foi realizado um experimento em um conjunto de registros bibliográficos. Essa abordagem parte de um pressuposto qualiquantitativo, buscando estabelecer uma relação dinâmica entre o mundo das análises dos dados e a teoria empregada, aplicando esta última para confrontar ou ratificar o que foi observado nos dados coletados.

O *corpus* do experimento constituiu-se de um conjunto de registros bibliográficos correspondentes aos artigos de periódicos selecionados por Souza (2006). O citado autor selecionou 60 artigos publicados em duas revistas da época: 29 artigos da Datagramazero

e 31 artigos da Ciência da Informação. Esses artigos foram escolhidos por sua importância e por sua qualidade reconhecida pelo Qualis da Capes naquele momento. Esse *corpus* está identificado na Tabela 5 do anexo A da tese de Souza (2005).

Ressalta-se que foi feito o reuso do *corpus* compilado por Silva, Correa e Gil-Leiva (2020), alterando os arquivos com extensão ".txt" para incluir, exclusivamente, os valores dos campos de metadados: título, resumo e as palavras-chave do autor. Os arquivos de extensão "key" foram mantidos inalterados, por conterem os termos da indexação intelectual controlada, realizada utilizando como linguagem de indexação o Tesauro Brasileiro de Ciência da Informação (PINHEIRO; FERREZ, 2014).

As etapas envolvidas no método proposto estão descritas na seção seguinte.

4 MÉTODO PROPOSTO

O método proposto consiste de duas etapas sequenciais, denominadas respectivamente: indexação automática e estudo métrico.

Para realizar a indexação automática por atribuição dos registros bibliográficos, foi aplicado o *software* Maui[4], tendo este sido escolhido por ser um software livre, adaptável para processar textos no idioma português do Brasil, e por realizar o processamento de texto livre, sem a necessidade de marcação de campos semânticos (como título, resumo e palavras-chaves). Adicionalmente, a indexação por atribuição automática empregada pelo Maui permite a utilização de um vocabulário controlado, que, para o método proposto, foi definido como sendo o Tesauro Brasileiro de Ciência da Informação (TBCI) (PINHEIRO; FERREZ, 2014).

Em Silva e Correa (2020) são descritas as técnicas de processamento de linguagem natural empregadas pelo software Maui, sendo indicada a leitura daquele artigo para os leitores que desejarem ter acesso a uma descrição mais detalhada do software. Foram seguidas as configurações do Maui reportadas no artigo de Silva, Correa e Gil-Leiva (2020), sendo a leitura desse último indicada por conter um relato de uso do Maui na indexação automática por atribuição do texto completo de artigos de periódicos brasileiros de Ciência da Informação.

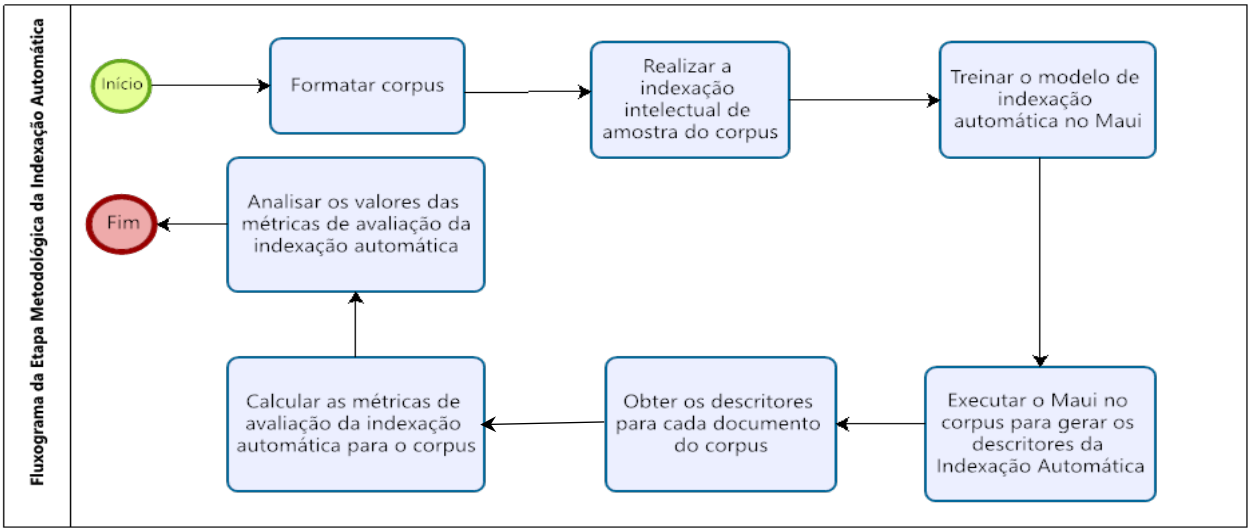
Para a realização de estudo métrico visando a formulação de indicadores temáticos, foram analisados gráficos gerados utilizando o *software* Iramuteq[5]. De acordo com Camargo e Justo (2013), o Iramuteq é um *software* livre que permite a viabilização de

[4] Disponível em: https://github.com/zelandiya/maui
[5] Disponível em: http://www.iramuteq.org/

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. Uma descrição mais detalhada do software Iramuteq pode ser encontrada no trabalho de Ferreira e Corrêa (2018).

Visando uma melhor descrição do método proposto, a Figura 1 apresenta os passos ou processos pertinentes ao fluxograma da etapa da indexação automática.

Figura 1 – Fluxograma da Indexação Automática



Fonte: dados da pesquisa (2021)

Cada passo da etapa de indexação automática é descrito a seguir:

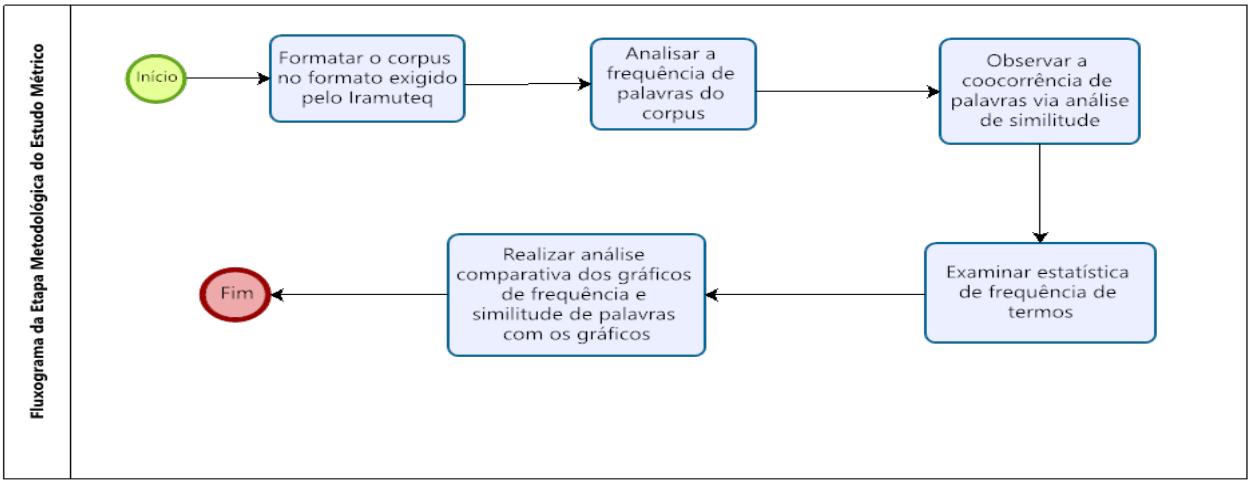
1. Formatar o *corpus* – Para cada registro bibliográfico do corpus, foi gerado um arquivo de texto de extensão ".txt", contendo os valores dos campos de metadados de título, resumo e palavras-chave, dispostos sequencialmente como texto sem formatação;
2. Realizar a indexação intelectual de amostra do *corpus* – Um subconjunto contendo 30 artigos foi indexado intelectualmente, usando o TBCI como linguagem de indexação, sendo os descritores incluídos em arquivos de texto de extensão "key". Foram reutilizados os arquivos de extensão "key" contendo o resultado da indexação intelectual dos 30 primeiros artigos do *corpus* (SILVA; CORREA; GIL-LEIVA, 2020). Tais arquivos são consultados no treinamento do modelo de indexação automática pelo Maui, sendo denominado como conjunto de treinamento;
3. Treinar o modelo de indexação automática no Maui – Realizou-se o treinamento de um modelo de indexação automática, via aprendizagem de máquina supervisionada no *software* Maui, para que este aprendesse a indexar automaticamente usando o TBCI. Nesse processo, foram utilizados os arquivos de texto contendo os metadados e termos da indexação intelectual relativos ao conjunto de treinamento. O modelo gerado pode

ser reutilizado na indexação automática de metadados de outro *corpus* da área da Ciência da Informação, tornando opcional as etapas 2 e 3;

4. Executar o Maui no *corpus* para gerar os descritores da indexação automática – O Maui foi executado para atribuir automaticamente os termos de indexação do TBCI para cada registro bibliográfico do *corpus*;
5. Obter os descritores para cada documento do *corpus* – Os descritores foram extraídos da saída do Maui para cada registro bibliográfico, obtendo-se o resultado da indexação automática realizada pelo sistema. Esse é o último passo dessa etapa, os demais passos de validação, a seguir, são opcionais;
6. Calcular as métricas de avaliação da indexação automática para o *corpus* – Visando à validação da indexação automática, foi gerado um relatório do *software* Maui contendo valores para métricas de qualidade da indexação automática. Tais métricas mensuram a semelhança dos descritores da indexação automática e da indexação intelectual. Foram reutilizados os arquivos de extensão "key" contendo o resultado da indexação intelectual dos 60 artigos do *corpus* (SILVA; CORREA; GIL-LEIVA, 2020) como padrão de referência ou padrão ouro. Para o cálculo dos índices, adotaram-se as métricas de consistência, revocação, precisão e medida F.
7. Analisar os valores das métricas de avaliação da indexação automática – A validação da indexação automática ocorreu com a análise dos valores das métricas calculadas.

A Figura 2 apresenta os passos ou processos pertinentes ao fluxograma da etapa de geração de indicadores temáticos, denominada de etapa do estudo métrico.

Figura 2 – Fluxograma do estudo métrico



Fonte: dados da pesquisa (2021)

Cada passo da etapa de estudo métrico é descrito a seguir:

1. Formatar o *corpus* no formato exigido pelo Iramuteq – O *corpus* foi formatado com os arquivos sendo alimentados com título, resumo e termos provenientes exclusivamente ou das palavras-chave, ou da indexação intelectual, ou da indexação automática. Para esse fim, foi realizada uma formatação de um arquivo de extensão ".txt", em que o conteúdo dos campos de metadados de cada artigo foi numerado com um padrão de caracteres (*** *Artigo_X, em que X corresponde a um inteiro único para cada artigo), e o arquivo foi salvo com codificação UTF-8 sem *Byte Order Mark* (BOM) no bloco de notas;
2. Analisar a frequência de palavras no *corpus* – Via Iramuteq, aplicou-se a Lei de Zipf no corpus, observando o grau de dispersão das palavras. Assim, foi possível comparar a frequência de ocorrência das palavras e verificar as mais frequentes.
3. Observar a coocorrência de palavras via análise de similitude – Foi utilizado o Iramuteq para gerar gráfico de análise de similitude para o corpus. Com o objetivo de alcançar uma melhor visualização dos agrupamentos temáticos de termos e obter os núcleos semânticos, foram aplicadas as configurações a seguir no Iramuteq: a) análise de similitude – Propriedades – Zerar as palavras-vazias; b) definição das palavras que seriam utilizadas nas análises; c) configurações gráficas visando destacar os agrupamentos e a deixar o tamanho do vértice proporcional à frequência. Nesse contexto, cabe ressaltar que os gráficos de similitude são visualizações que auxiliam na compreensão semântica e temática, possibilitando identificar as coocorrências entre as palavras, trazendo indicações de conexidade e contribuindo na identificação da estrutura de um *corpus* textual (FERREIRA; CORRÊA, 2018).
4. Examinar estatística de frequência de termos – Gráficos de frequência de ocorrência no *corpus* foram construídos e analisados para palavras-chave, descritores da indexação intelectual e descritores da indexação automática, respectivamente. Os dois primeiros, para fins de validação do último;
5. Realizar análise comparativa dos gráficos de frequência e similitude de palavras com os gráficos de frequência de termos – Analisaram-se os gráficos, contrapondo cada um dos agrupamentos temáticos de palavras e os termos frequentes, a fim de atestar a validade do método proposto e apontar os indicadores temáticos do *corpus*.

5 ANÁLISE DE RESULTADOS

Neste estudo, utilizou-se como *corpus* um conjunto de registros bibliográficos descritivos dos 60 artigos de periódicos selecionados por Souza (2005) em sua tese de doutorado. Ressalta-se que foi feito o reuso, com adaptações, do *corpus* como compilado por Silva, Correa e Gil-Leiva (2020), alterando os arquivos de extensão ".txt" para incluir somente o texto dos campos de metadados: título, resumo e as palavras-chave do autor. Os arquivos de extensão ".key", por sua vez, não foram alterados, sendo mantida a indexação intelectual.

A escolha desse conjunto de documentos foi definida pelo fato de estes já terem sido analisados por outros autores da área de Ciência da Informação. Como o propósito deste artigo é validar um método proposto, o reuso desses dados possibilita observar, de forma criteriosa, os resultados obtidos.

5.1 Análise da indexação automática

Segundo Narukawa, Gil-Leiva e Fujita (2009) a avaliação da qualidade da indexação pode ser feita via análise da consistência na indexação, que reflete o grau de concordância na representação da informação de um documento por diferentes indexadores. Adicionalmente, a avaliação extrínseca da indexação automática pode ser realizada por meio de índices de revocação, precisão e medida F, tendo a indexação intelectual como padrão de referência (SILVA; CORREA; GIL-LEIVA, 2020).

Portanto, os termos obtidos por meio da indexação automática, atribuídos pelo *software* Maui, foram comparados aos termos da indexação intelectual, sendo calculados o número de termos comuns e os índices de consistência, precisão, revocação e medida F na indexação automática para cada documento do *corpus*. Os resultados foram dispostos em uma tabela, contemplando os seguintes campos: número de termos extraídos pelo Maui; número de termos indexados manualmente; número de termos comuns entre as duas indexações; consistência; precisão; revocação; e medida F.

Os parâmetros descritivos dos valores das métricas de avaliação da indexação automática para o *corpus* estão sintetizados na Tabela 1, a seguir.

Tabela 1: Valores das métricas obtidas pelo Maui no *corpus*

	Número de termos do Maui	Número de termos da indexação intelectual	Número de termos comuns com a indexação intelectual	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	2,08	1,58	13,6%	18,4%	17,4%	15,2%

Fonte: dados da pesquisa (2021)

A partir desses resultados, observou-se um mínimo de 7 palavras indexadas e um máximo de 15 termos atribuídos pela indexação intelectual. Enquanto o Maui atribuiu automaticamente um mínimo de 4 termos e um máximo de 10. Esses valores resultam em uma média de 10 descritores atribuídos por documento pela indexação intelectual, de um total de 621, e uma média de 8 descritores atribuídos automaticamente por documento, de um total de 495, pelo Maui.

Outro ponto relevante diz respeito à consistência, percebe-se uma média de 33% nos índices de consistência da indexação automática, variando entre 11% e 54%. A consistência média obtida pode ser categorizada como um nível de desempenho bom na indexação automática, segundo a categorização proposta por Bandim e Correa (2018).

Com relação à precisão, foi identificada uma média de 56%, e, para a revocação, uma média de 45%, indicando um desempenho esperado para o conjunto analisado. Já em relação à medida F, foi obtido um índice médio de 48%. Por ser uma média harmônica entre o índice de precisão e o de revocação, a medida F identifica o percentual médio de termos relevantes recuperados da indexação intelectual. Nesse caso, se nenhum termo relevante fosse recuperado, seria considerado o valor de zero; se todos os termos recuperados fossem relevantes, assumir-se-ia o valor de 1 ou 100%. Dada a média de 48%, pode-se afirmar que quase metade dos termos atribuídos ou recuperados pela indexação automática são relevantes, isto é, são termos da indexação intelectual. Embora os resultados não sejam diretamente comparáveis, os valores médios para as métricas de avaliação da indexação automática são superiores aos reportados por Narukawa, Gil-Leiva e Fujita (2009).

Bandim e Correa (2019) fizeram uso do mesmo conjunto de documentos, porém utilizaram o sistema SISA na indexação automática do texto completo dos artigos e fizeram uso das palavras-chave do autor como padrão de referência. Os valores médios para as

métricas de avaliação da indexação automática reportados no presente trabalho são superiores aos encontrados no trabalho dos referidos autores.

Ao serem analisados os resultados apontados por Silva, Correa e Gil-Leiva (2020), identificam-se valores médios das métricas próximos aos obtidos neste trabalho. Esses autores utilizaram o Maui na indexação automática do texto completo dos artigos e usaram os termos da indexação intelectual como padrão de referência. O diferencial do presente trabalho está no uso dos metadados no lugar do texto completo como entrada para o sistema de indexação automática.

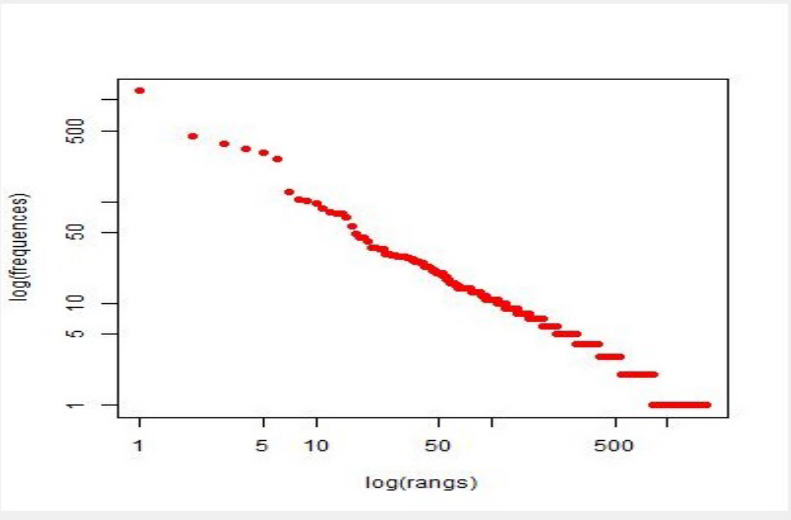
Na investigação dos supracitados pesquisadores, o índice de consistência médio obtido na indexação automática pelo Maui foi de 36%, próximo ao índice de 33% encontrado na Tabela 1. A precisão média foi de 56% no presente trabalho, próximo aos 54% reportados pelos autores supracitados. A mesma proximidade é identificada no índice de revocação média, com 45% ante os 51% reportados pelos autores supracitados. O índice de medida F média apresentou-se em 48% ante os 52% reportados pelos autores supracitados. Nessa circunstância, confirma-se que os valores médios dos índices extraídos em ambos os trabalhos são próximos.

O processo de avaliação da indexação automática aponta para bons resultados na representação temática dos registros bibliográficos. Caso os resultados não se enquadrassem no padrão estabelecido como bom, a indexação intelectual, a codificação da tabela de caracteres dos arquivos de entrada, o uso do tesauro nas indexações e as ferramentas de processamento de linguagem natural do sistema de indexação automática precisariam ser revisados, com o propósito de atingir bons índices de consistência, precisão, revocação e medida F.

5.2 Análise estatística de palavras

A primeira análise estatística foi mensurada por meio do *software* Iramuteq na aplicação da Lei de Zipf, observando o grau de dispersão das palavras no *corpus*. Para sua elaboração, foi levado em consideração o texto dos metadados (título, resumo e palavras-chave), o que resultou no Gráfico 1, a seguir, o qual apresenta a classificação e frequência das palavras nos metadados em escala logarítmica.

Gráfico 1 – Frequência das palavras nos metadados (título, resumo e palavras-chave do autor)



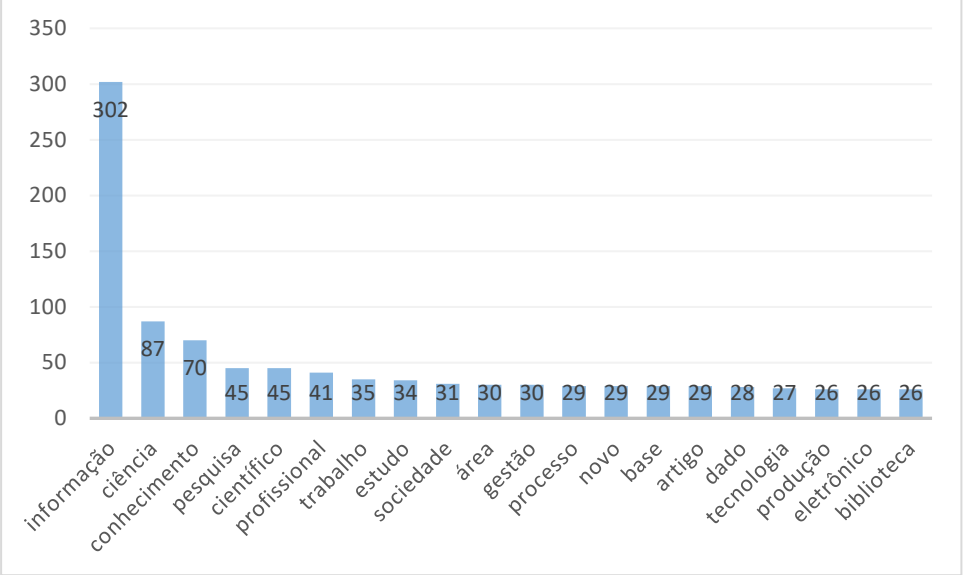
Fonte: dados da pesquisa (2021)

A aplicação da Lei de Zipf no *corpus*, de acordo com Alvarado (1984), demonstra níveis de frequência das palavras. Além disso, Araújo (2006) pondera que a lei discrimina a relação existente entre as palavras e seu uso. Assim, verifica-se uma correlação entre o número de palavras diferentes e a frequência de seu uso, existindo uma regularidade na seleção e no emprego das palavras, em que um pequeno número de palavras ocorrem mais frequentemente e um grande número de palavras ocorrem menos vezes.

No Gráfico 1, para os 60 registros bibliográficos, foram identificadas 9053 ocorrências de palavras para 1644 formas de palavras (lemas de palavras significativas). Entretanto, cerca de 9,11% das ocorrências são de 825 formas que aparecem apenas uma vez nos documentos, e que representam 50,18% das formas de palavras. Sendo assim, pode-se afirmar que pelo menos metade das formas de palavras ocorrem apenas uma vez, sugerindo uma configuração de dispersão, tendo em vista que a outra metade das formas de palavras correspondem a 89,89% das ocorrências. Ressalta-se que nessa análise, as palavras vazias de significado (*stopwords*) foram desconsideradas, e cada palavra restante (ou significativa) foi representada por sua forma canônica ou lema, sendo considerada isoladamente. Os termos compostos serão abordados posteriormente.

Percebe-se, portanto, a tendência de repetição de algumas palavras, o que pode ser comprovado pela distribuição de frequência absoluta apresentada no Gráfico 2.

Gráfico 2 – Frequência das 20 formas de palavras mais frequentes no *corpus*



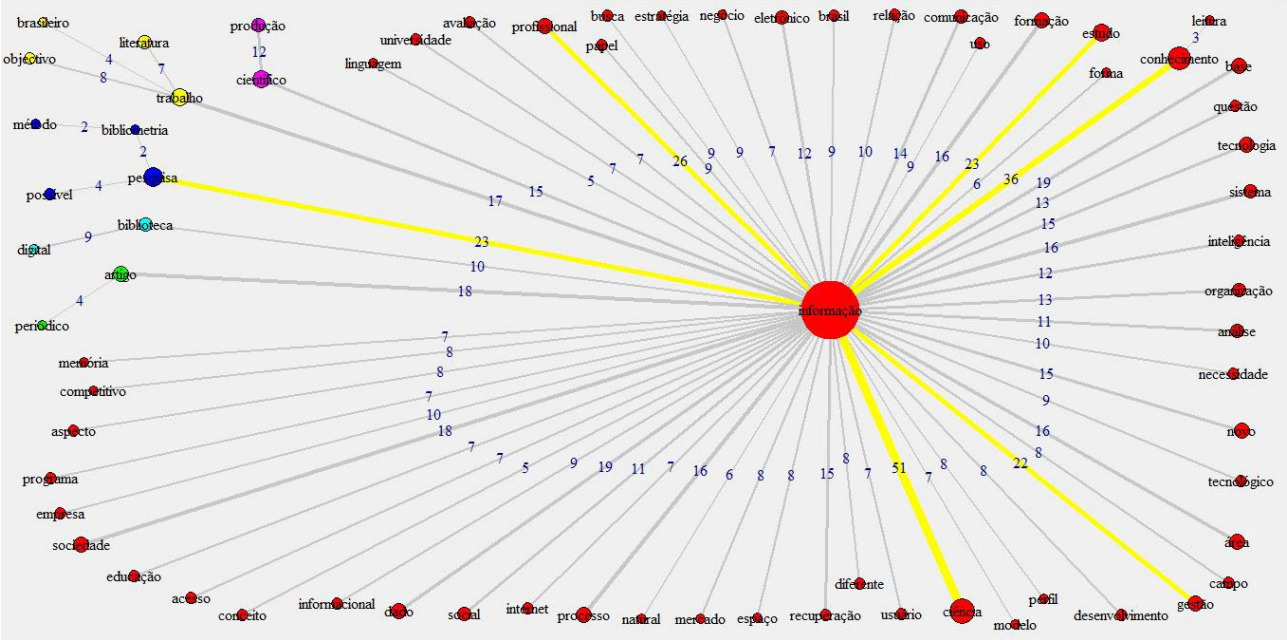
Fonte: dados da pesquisa (2021)

O Gráfico 2 comprova a aplicação da Lei de Zipf ao identificar um pequeno conjunto de palavras com alta ocorrência no *corpus*. Constatou-se que as palavras “informação”, “ciência” e “conhecimento” são os termos dominantes, evidenciando o núcleo temático do *corpus* estudado. Adicionalmente, as palavras “pesquisa”, “científico” e “profissional” encontram-se entre as seis mais frequentes do *corpus* em destaque no topo do Gráfico 1.

Posteriormente, realizou-se a análise de similitude no método proposto, visando à identificação de termos da linguagem de especialidade. Tal análise decorre do agrupamento das palavras, as quais se unem por critérios relacionais e de similitude em um grafo, baseando-se na teoria dos grafos, frequentemente adotada por pesquisadores que trabalham com representações sociais. Esse agrupamento possibilita a identificação das coocorrências entre as palavras, trazendo indicações do grau de conexidade existente e auxiliando na identificação de termos, pois representa a proximidade sintática entre as palavras na formação de termos compostos no *corpus*.

O Gráfico 3 foi elaborado a partir dos metadados, com as 70 palavras mais frequentes no *corpus*, onde a ocorrência varia de 11 a 302.

Gráfico 3 – Análise de agrupamentos de palavras no *corpus* a partir dos metadados (título, resumo e palavras-chave do autor)



Fonte: dados da pesquisa (2021)

Nesse tipo de visualização, é possível identificar um polo central referente à palavra “informação”, que possui diversas palavras tematicamente conectadas em seu entorno. Na análise quantitativa, verificou-se que “informação” é a palavra mais frequente, apresentando 302 ocorrências.

Observa-se que as palavras se ligam a outras por meio de linhas nas cores amarela ou cinza-claro. Na teoria dos grafos, essas linhas são denominadas arestas, cuja espessura representa o grau de conexão entre as palavras. Nesse caso, percebe-se que “informação” se conecta principalmente com as palavras “ciência”, “conhecimento”, “profissional”, “pesquisa”, “estudo” e “gestão”.

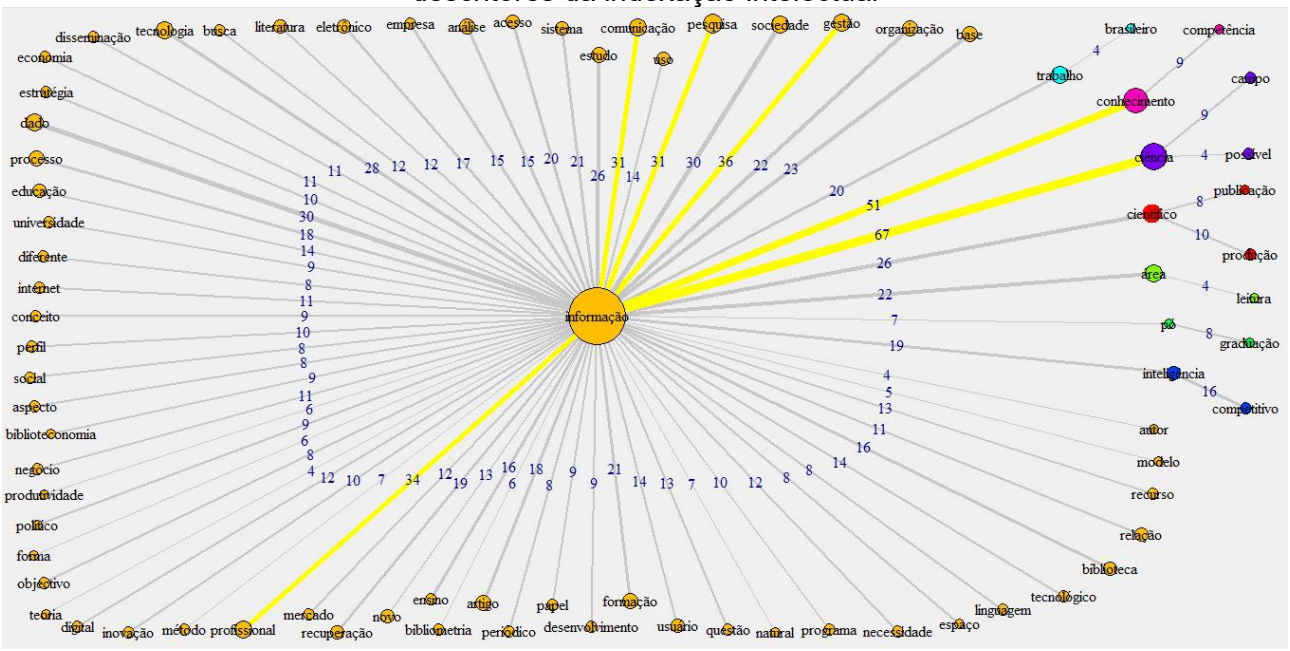
A palavra “informação” está mais intensamente conectada à palavra “ciência”, isso ocorre em razão do domínio do *corpus* estudado, pois os 60 artigos pertencem ao campo da “Ciência da Informação”. Dentre os elementos desse campo, são estudadas relações fenomenológicas, como a relação entre “informação e conhecimento”. Por esse motivo, a palavra “conhecimento” conecta-se com “informação”, apresentando um índice de força de ocorrência no valor de 36. A palavra “informação” conecta-se também às palavras “profissional” e “gestão”, correspondendo respectivamente, aos termos “profissional da informação” e “gestão da informação”.

Algumas palavras aparecem em agrupamentos externos ao primeiro agrupamento central, com nodos na cor vermelha, que contém a palavra “informação” como núcleo.

Dentre os agrupamentos externos, destacam-se os que sugerem termos de especialidade: o agrupamento na cor azul-escura tem a palavra “pesquisa” conectada à “bibliometria”, o que pode decorrer do emprego da bibliometria como método de pesquisa por alguns artigos; o agrupamento na cor verde apresenta como núcleo a palavra “artigo” conectada à palavra “periódico”, sugerindo o “artigo de periódico” como objeto principal das análises; o agrupamento na cor rosa traz as palavras “científico” e “produção”, que decorrem do termo “produção científica”; por fim, o agrupamento na cor azul-claro apresenta as palavras “biblioteca” e “digital”, as quais são comumente utilizadas no termo “biblioteca digital”.

Nos Gráficos 4 e 5, é possível observar os principais agrupamentos de palavras extraídos dos metadados com as palavras-chave do autor substituídas, respectivamente, pelos descritores atribuídos pela indexação intelectual e pela indexação automática.

Gráfico 4 – Análise de agrupamentos de palavras do *corpus* a partir do título, do resumo e dos descritores da indexação intelectual

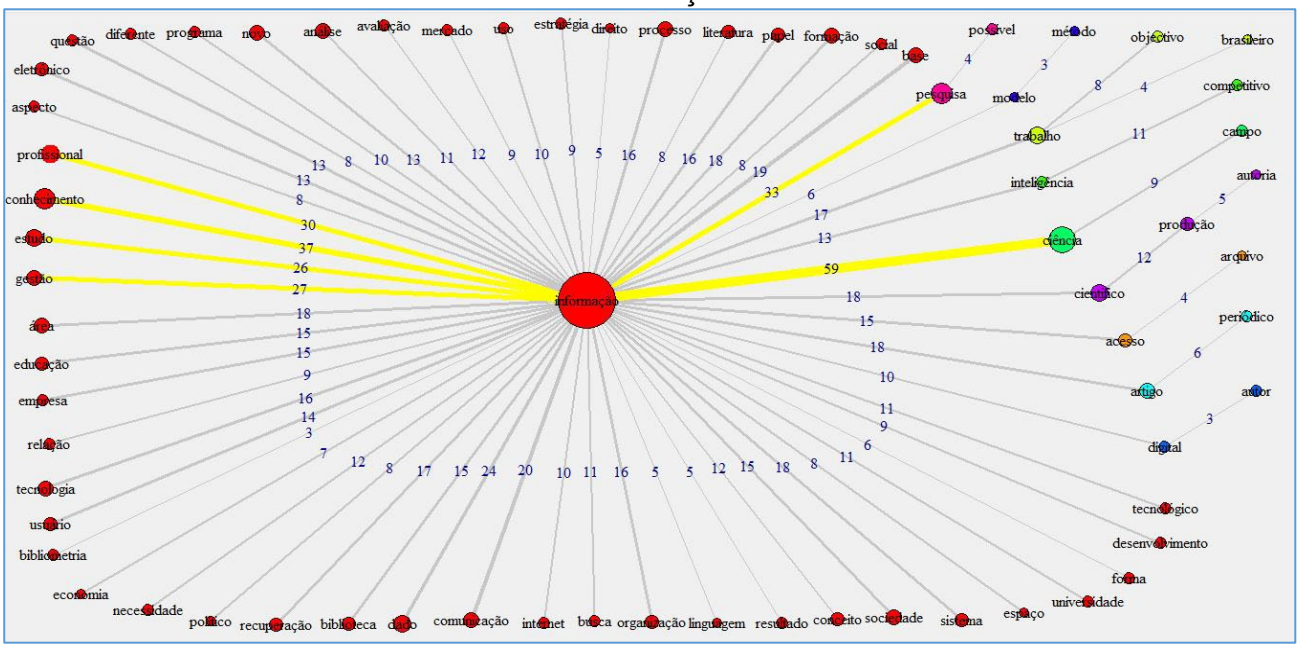


Fonte: dados da pesquisa (2021)

O Gráfico 4 apresenta um agrupamento central na cor amarela, tendo como núcleo a palavra “informação”, que se conecta com a maioria dos termos observados, sendo vinculada mais fortemente às palavras “ciência”, “conhecimento”, “gestão”, “profissional”, “comunicação” e “pesquisa”. Dentre os possíveis termos de especialidade que emergem das relações entre as palavras mais coocorrentes desse agrupamento, tem-se: “ciência da informação”, “gestão da informação” e “profissional da informação”.

Analisando os demais agrupamentos no Gráfico 4, identificam-se relacionados a termos de especialidade os seguintes: o agrupamento na cor azul-escuro, que apresenta as palavras “inteligência” e “competitivo”, tornando possível estabelecer a relação com o termo “inteligência competitiva”; o agrupamento na cor lilás, representado pelas palavras “científico”, “publicação” e “produção”, sendo comumente utilizado nos termos “produção científica” e “publicação científica”; e o agrupamento, na cor verde-escura, das palavras “pós” e “graduação”, relacionado ao termo “pós-graduação”.

Gráfico 5 – Análise de agrupamentos de palavras do *corpus* a partir do título, do resumo e dos descritores da indexação automática



Fonte: dados da pesquisa (2021)

No Gráfico 5, é possível identificar um agrupamento central na cor vermelha, com o termo “informação” como núcleo, conectando-se mais fortemente aos termos “ciência”, “pesquisa”, “conhecimento”, “profissional”, “gestão” e “estudo”. Os termos de especialidade que emergem das relações entre as palavras mais coocorrentes desse agrupamento são “ciência da informação”, “profissional da informação” e “gestão da informação”.

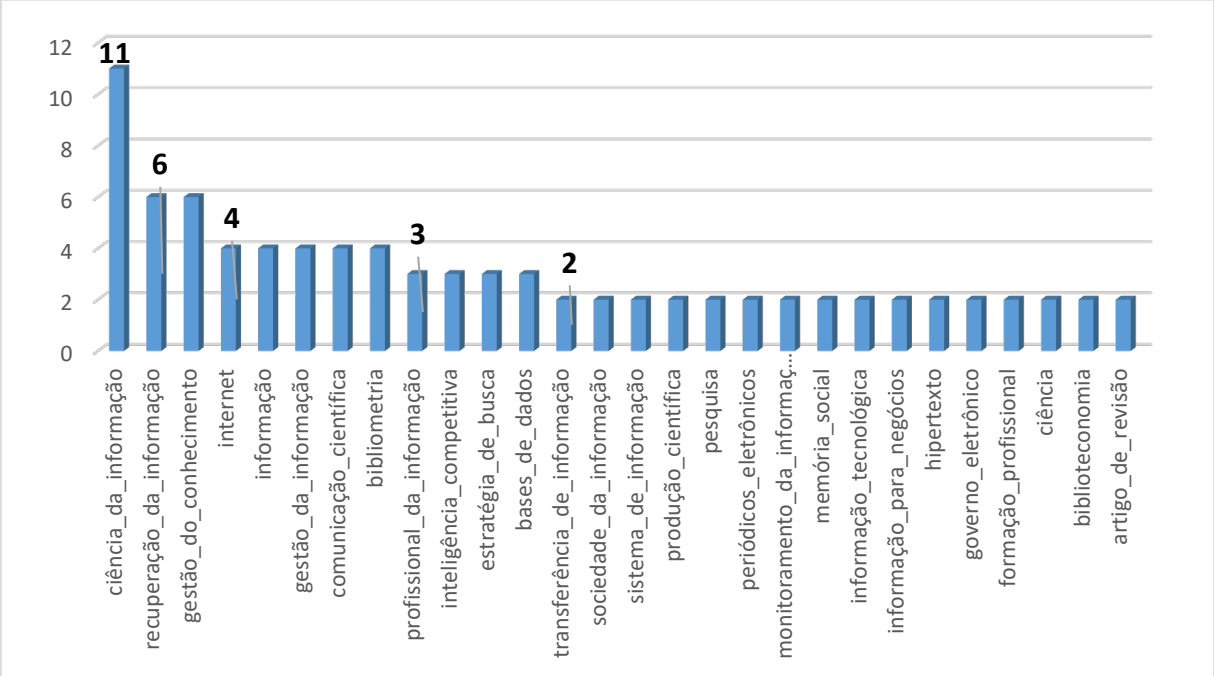
Existem também agrupamentos externos com coocorrência de palavras menos frequentes. Dentre os agrupamentos externos, mostram-se significativos os seguintes: entre as palavras “inteligência” e “competitivo”, na cor verde, representando o termo “inteligência competitiva”; entre as palavras “científico” e “produção”, na cor roxa, representando o termo “produção científica”; e entre as palavras “artigo” e “periódico”, na cor azul-claro, representando o termo “artigo de periódico”.

Assim, percebem-se semelhanças entre os três gráficos de similitude, como as palavras mais frequentes em comum e a preservação das relações mais fortes de coocorrência entre palavras, sendo algumas dessas relações derivadas de termos compostos da linguagem de especialidade. Destarte, pode-se concluir que a análise estatística de palavras isoladas, levando em conta os metadados de título, resumo e palavras-chave do autor, fornece parâmetros para validar a análise estatística de palavras e de termos atribuídos pela indexação automática.

5.3 Análise estatística de termos

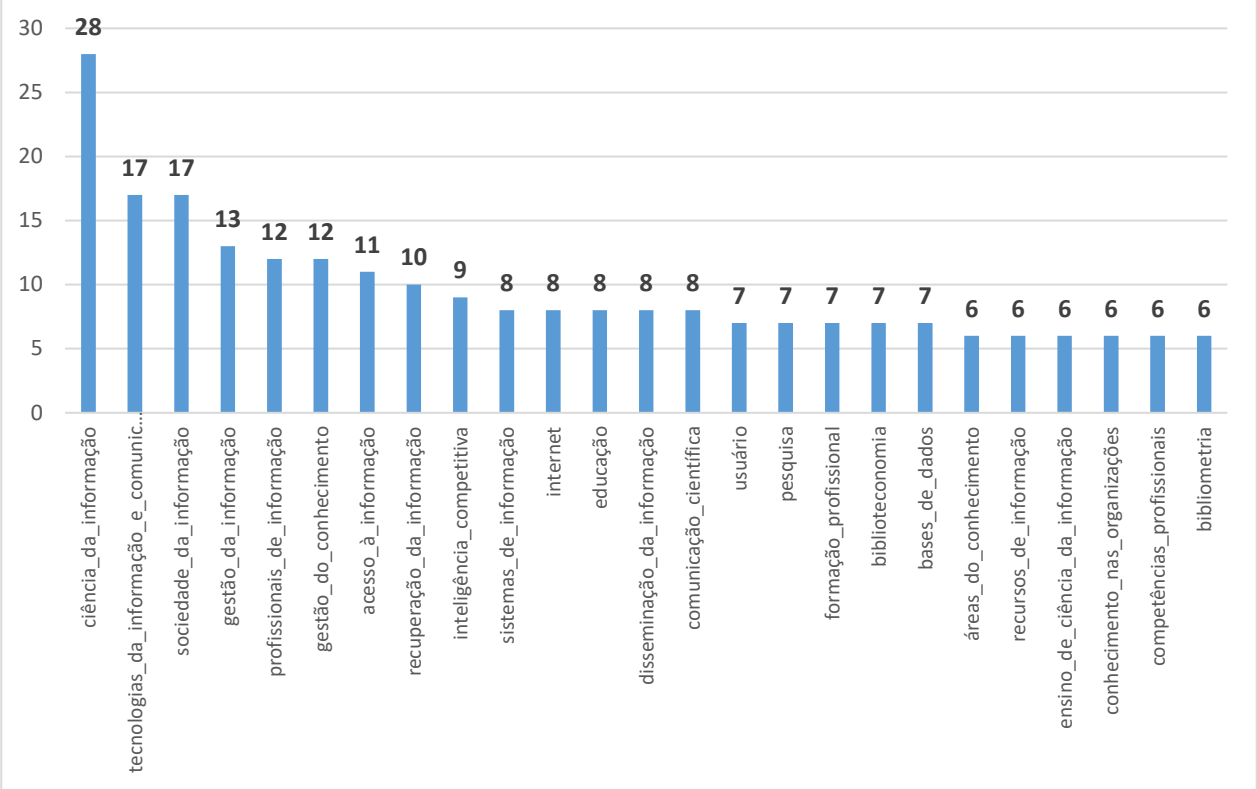
Os Gráficos 6 e 7, adiante expostos, apresentam a frequência absoluta dos termos mais recorrentes, respectivamente, nas palavras-chave do autor e nos descritores da indexação intelectual. Devido à dispersão terminológica, o Gráfico 6 apresenta uma distribuição de frequência achatada e de cauda longa, com somente três termos ocorrendo como assunto em 10% ou mais dos artigos do *corpus*: “ciência da informação”, “recuperação da informação” e “gestão do conhecimento”.

Gráfico 6 – Frequência de palavras-chave do autor mais frequentes



Fonte: dados da pesquisa (2021)

Gráfico 7 – Os 25 descritores mais frequentes da indexação intelectual no *corpus*



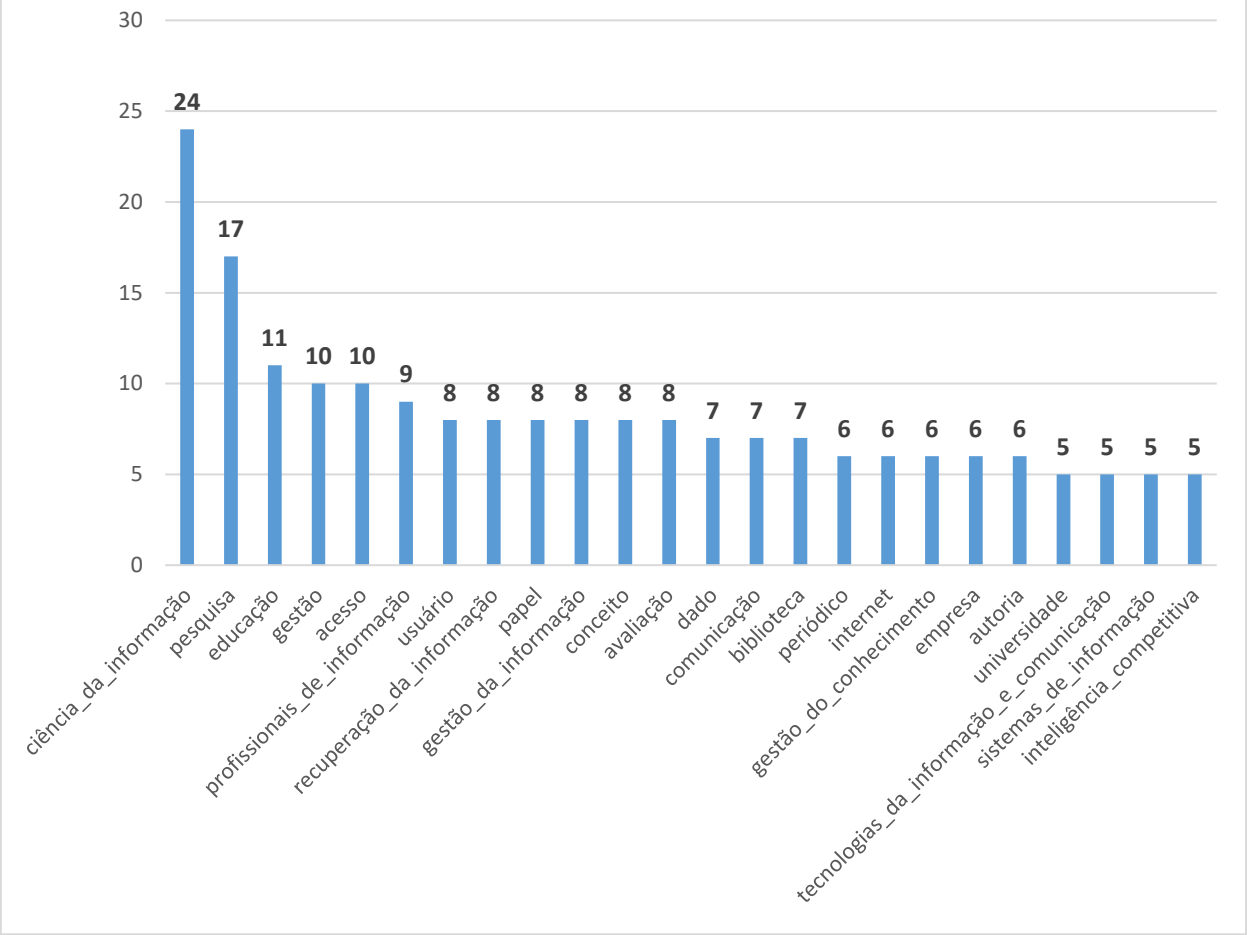
Fonte: dados da pesquisa (2021)

No Gráfico 7, nota-se uma distribuição de frequência com maiores valores de ocorrência, sendo “ciência da informação”, “tecnologias da informação e comunicação”, “sociedade da informação”, “gestão da informação”, “profissionais da informação” e “gestão do conhecimento” os descritores mais frequentes. Cerca de 25 descritores ocorrem em 10% ou mais dos artigos do *corpus* e incluem os termos mais frequentes das palavras-chave do autor.

Entretanto, no mesmo gráfico, evidenciam-se descritores frequentes não contemplados nas palavras-chave mais recorrentes, como “tecnologias da informação e comunicação”, “acesso à informação”, “educação”, “disseminação da informação” e “usuário”. Em contrapartida, a palavra-chave “informação”, presente no Gráfico 6, não aparece no Gráfico 7, pois não se configura como um descritor no TBCI.

O Gráfico 8, a seguir, apresenta a distribuição de frequência absoluta dos descritores mais frequentemente atribuídos pela indexação automática.

Gráfico 8 – Frequência de descritores do Maui para o *corpus*



Fonte: dados da pesquisa (2021)

Os três gráficos são validados pela análise dos gráficos de similitude e do gráfico de frequência das palavras isoladas. Ao observar os termos em sua configuração composta, compreende-se que a palavra “informação” se encontra presente em vários termos compostos que se repetem no *corpus* analisado. Do mesmo modo, a palavra “ciência” tem seu significado mais frequentemente atrelado ao descritor “ciência da informação”. A palavra “conhecimento”, em seu turno, tem seu significado majoritariamente vinculado ao descritor “gestão do conhecimento”. Assim, observa-se que os termos compostos mais recorrentes caracterizam bem os assuntos tratados nos documentos do *corpus*.

Na comparação entre os três gráficos, nota-se uma variação na distribuição de frequência de termos mais frequentes. É perceptível uma mudança tanto quantitativa, na frequência, quanto na atribuição de determinados termos. Nesse sentido, o Maui, por meio do método de aprendizagem de máquina e do uso de vocabulário controlado, atribuiu mais frequentemente que o autor dos trabalhos os seguintes termos: “ciência da informação”, por 13 unidades a mais; “recuperação da informação”, por duas unidades a mais; e “gestão da informação”, por quatro unidades a mais. Os descritores “gestão do conhecimento” e

“comunicação científica” não tiveram aumento de frequência de atribuição pelo Maui. Para facilitar a comparação de tais resultados, veja a Tabela 2.

Tabela 2 **–** Análise comparativa da frequência dos termos compostos mais frequentes no *corpus*

	Ciência da Informação	Recuperação da Informação	Gestão do Conhecimento	Gestão da Informação	Comunicação Científica	Profissionais de Informação
Autor	11	6	6	4	4	3
Indexador	28	10	12	13	8	12
Maui	**24**	**8**	**6**	**8**	**4**	**9**
Variação-A	*+13*	*+2*	*0*	*+4*	*0*	*+6*
Variação-I	*-4*	*-2*	*-6*	*-5*	*-4*	*-3*

Fonte: dados da pesquisa (2021)

A partir da Tabela 2, é possível observar que há um conjunto expressivo de termos compostos mais frequentes em comum nas três indexações, porém as frequências de atribuição são diferentes para cada termo em cada tipo de indexação. Para os termos compostos mais frequentes em comum com a indexação do autor, o Maui gera uma distribuição de frequência com variação positiva (Variação-A na Tabela 2), isto é, com termos com frequência igual ou maior que as palavras-chaves do autor. Todavia, o Maui gera uma distribuição de frequência com variação negativa para esses termos na indexação intelectual (Variação-I na Tabela 2). Isso mostra que a indexação automática por atribuição comporta-se gerando uma distribuição meio-termo de frequência de termos compostos, entre aquelas obtidas via palavras-chave do autor e termos da indexação intelectual.

Quanto aos unitermos, o Maui priorizou a indexação dos seguintes termos mais frequentes: “pesquisa”, “gestão”, “educação”, “acesso”, “conceito”, “usuário”, “papel” e “avaliação”. Dentre esses, são comuns com a indexação intelectual: “pesquisa”, “educação” e “usuário”.

Nas palavras-chaves do autor, os unitermos “internet” e “bibliometria” aparecem entre os mais frequentes, sendo tais termos também presentes entre os mais frequentes nas outras duas indexações. Adicionalmente, o termo “pesquisa” aparece com frequência igual a dois nas palavras-chave do autor, com frequência igual a sete na indexação intelectual e 17 na indexação automática.

O Maui também propôs unitermos que não aparecem entre os mais frequentemente atribuídos pelos indexadores no *corpus*, como: “gestão”, “acesso”, “conceito”, “papel” e “avaliação”. Diante disso, entende-se ser necessária uma revisão mais cuidadosa dos unitermos mais frequentes, principalmente para os que não constam como palavras-chave do autor mais recorrentes.

Nesse contexto, um procedimento viável para a geração de indicadores temáticos consiste na seleção dos descritores mais frequentes comuns entre as palavras-chave do

autor e a indexação automática como os termos mais representativos do *corpus*. Isso é possível porque a indexação automática, ao realizar uma indexação mais uniforme, aponta para uma frequência de tais termos mais próxima do ideal da indexação intelectual.

Com base nos resultados obtidos na aplicação do método proposto ao *corpus*, elaboraram-se duas inferências principais sobre os resultados alcançados nesta análise:

a) 1.ª inferência: A produção científica do *corpus* apresentou enfoque em assuntos como: Ciência da Informação, profissionais da informação, recuperação da informação, gestão da informação, gestão do conhecimento e comunicação científica.
b) 2.ª inferência: Constatou-se a eficácia do método proposto divido em etapas, primeiramente com a adoção dos aportes da indexação automática e, depois, dos aportes dos estudos métricos da informação, permitindo a obtenção de indicadores temáticos do *corpus* analisado.

6 CONSIDERAÇÕES FINAIS

Os resultados apresentados apontam que os indicadores temáticos gerados pela aplicação do método proposto representam os principais temas identificados no *corpus*. Verificou-se, também, que o método pode ser aplicado na elaboração de indicadores temáticos de outros conjuntos de registros bibliográficos da área de Ciência da Informação, sem a necessidade de indexação intelectual de uma amostra dos registros por parte do pesquisador, dada a possibilidade de reuso dos dados da presente pesquisa.

Ademais, foi possível constatar que o método proposto para construção de indicadores temáticos de informação científica na área de Ciência da Informação, contribui para a sistematização do processo de construção ao aplicar conjuntamente as técnicas da indexação automática e dos estudos métricos da informação, permitindo a diminuição da sobrecarga intelectual do pesquisador na realização de estudos de mapeamento temático e mapeamento da ciência.

Ao adotar o processo de indexação automática por atribuição, usando o TBCI como linguagem de indexação, foram alcançados os principais termos relevantes, que, validados pelas análises bibliométricas de palavras isoladas e de palavras-chave do autor, permitem segurança na delimitação de indicadores temáticos do *corpus* analisado.

O método proposto permite acelerar o processo de descoberta de conhecimento, por permitir a extração automatizada dos conceitos principais abordados em um determinado conjunto de registros bibliográficos, possibilitando maior agilidade na identificação de

descritores relevantes por meio de análises bibliométricas, e, consequentemente, maior agilidade na obtenção de indicadores temáticos.

Como limitações do método proposto, podem-se apontar: a dependência da atualização do tesauro para estudos que analisem registros publicados cinco anos após a publicação do mesmo, que pode não incluir termos consolidados mais recentemente; a dependência a um modelo de indexação automática com boa eficácia; e a possibilidade de o modelo de indexação automática cometer erros na atribuição automática de descritores aos registros bibliográficos.

Estudos futuros propõem ampliar as análises obtidas por meio da aplicação de outros sistemas de indexação automática por atribuição, como o SISA e o KEA. Além disso, pretende-se aplicar o método proposto em um *corpus* diferente de registros bibliográficos de publicações da área de Ciência da Informação no Brasil e, depois, de outras áreas do conhecimento. Adicionalmente, almeja-se utilizar o *software VOSviewer* para a análise de agrupamento e de coocorrência de descritores, com o propósito de gerar mapas temáticos.

REFERÊNCIAS

ARAÚJO, C. A. A. Bibliometria: evolução histórica e questões atuais. **Em Questão**, Porto Alegre, v. 12, n. 1, p. 11-32, jan./jun. 2006. Disponível em: https://seer.ufrgs.br/index.php/EmQuestao/article/view/16. Acesso em: 19 maio 2023.

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

ALVARADO, R. U. Bibliometria no Brasil. **Ciência da Informação**, Brasília, v.13, n.2, p. 91-105, jul./dez. 1984. Disponível em: https://revista.ibict.br/ciinf/article/view/200. Acesso em: 19 maio 2023.

ALVARADO, R. U. A Bibliometria: história, legitimação e estrutura. *In*: TOUTAIN, L. M. B. B. (org.). **Para entender a ciência da informação**. Salvador: EDUFBA, 2007. p. 185-217.

BANDIM, M. A. S.; CORREA, R. F. A consistência na indexação automática por atribuição de artigos científicos na área de Ciência da Informação. **Encontros Bibli**: revista eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, v. 23, n. 53, p. 64-77, 2018. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23n53p64. Acesso em: 19 maio 2023.

BANDIM, M. A. S.; CORREA, R. F. Indexação automática por atribuição de artigos científicos em português da área de Ciência da Informação. **Transinformação**, Campinas, v. 31, 2019. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-37862019000100501&lng=en&nrm=iso. Acesso em: 19 maio 2023.

BEIRA, J. C. *et al*. Indicadores bibliométricos na produção científica em periódicos brasileiros da Ciência da Informação no estrato A1. **Revista ACB**: Biblioteconomia em Santa Catarina, Florianópolis, v. 25, n. 2, p. 383-408, abr./jul., 2020. Disponível em: https://revista.acbsc.org.br/racb/article/view/1660. Acesso em: 19 maio 2023.

BORNER, K.; CHEN, C.; BOYACK, K. W. Visualizing knowledge domains. **Annual Review of Information Science and Technology**, [*S.l.*] v. 37, n.1, p. 179-255, 2003. Disponível em: https://asistdl.onlinelibrary.wiley.com/doi/10.1002/aris.1440370106. Acesso em: 19 maio 2023.

BHUYAN, A.; SANGURI, K.; SHARMA, H. Improving the Keyword Co-occurrence Analysis: An Integrated Semantic Similarity Approach. *In*: IEEE INTERNATIONAL CONFERENCE ON INDUSTRIAL ENGINEERING AND ENGINEERING MANAGEMENT (IEEM), 2021, Singapore. **Proceedings of […]**, Singapore: IEEE, 2021. p. 482-487. Disponível em: https://ieeexplore.ieee.org/document/9673030. Acesso em: 19 maio 2023. DOI: 10.1109/IEEM50564.2021.9673030.

CAMARGO, B. V; JUSTO, A. M. IRAMUTEQ: um software gratuito para análise de dados textuais. **Temas em Psicologia**, Ribeirão Preto, v. 21, n. 2, p. 513-518, 2013. Disponível em: http://pepsic.bvsalud.org/scielo.php?pid=S1413-389X2013000200016&script=sci_abstract. Acesso em: 19 maio 2023.

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. Disponível em: https://periodicos.ufpb.br/index.php/itec/article/view/21408. Acesso em: 19 maio 2023.

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**, [*S.l.*], v. 25, n. 1, p. 67-78, 1999. Disponível em: https://journals.sagepub.com/doi/10.1177/016555159902500107. Acesso em: 19 maio 2023.

DOLOREUX, D. *et al*. Territorial innovation models: to be or not to be, that's the question. **Scientometrics,** [*S.l.*], v.120, p. 1163–1191, 2019. Disponível em: https://link.springer.com/article/10.1007/s11192-019-03181-1. Acesso em: 19 maio 2023.

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: 19 maio 2023.

GIL-LEIVA, I.; ORTUÑO, P. D.; CORRÊA, R. F. Indización automática de artículos científicos sobre Biblioteconomía y Documentación con SISA, KEA y MAUI. **Revista Española de Documentación Científica**, [*S. l.*], v. 45, n. 4, p. e338, 2022. Disponível em: https://redc.revistas.csic.es/index.php/redc/article/view/1371. Acesso em: 19 maio 2023.

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: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-37862006000100003&lng=en&nrm=iso. Acesso em: 19 maio 2023.

KOBASHI, N. Y.; SANTOS, R. N. M. D. Arqueologia do trabalho imaterial: uma aplicação bibliométrica à análise de dissertações e teses. **Encontros Bibli**: revista eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, n. esp. 1. sem., p. 106-115, 2008. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2008v13nesp1p106. Acesso em: 19 maio 2023.

KOBASHI, N. Y.; DÍAZ, F.; SANTANA, S. Cartografia temática e de colaboração em organização do conhecimento no Brasil (2000-2010). **Ciência da Informação**, Brasília , v. 43, n. 1, 2014. Disponível em: https://revista.ibict.br/ciinf/article/view/1417. Acesso em: 19 maio 2023.

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

NARUKAWA, C. M.; GIL-LEIVA, I.; FUJITA, M. S. L. Indexação automatizada de artigos de periódicos científicos: análise da aplicação do *software* SISA com uso da terminologia DeCS na área de odontologia. **Informação & Sociedade**: Estudos, João Pessoa, v. 19, n. 2, p. 99-118, 2009. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/2925. Acesso em: 19 maio 2023.

NORONHA, D. P.; MARICATO, J. M. Estudos métricos da informação: primeiras aproximações. **Encontros Bibli**: revista eletrônica de Biblioteconomia e Ciência da Informação, Florianópolis, v. 13, n. 1, p. 116-128, 2008. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2008v13nesp1p116. Acesso em: 19 maio 2023.

NOYONS, E.; VAN RAAN, A. Bibliometric cartography of scientific and technological developments of an R & D field: the case of optomechatronics. **Scientometrics**, [*S.l.*], v. 30, n. 1, p. 157-173, 1994. Disponível em: https://link.springer.com/article/10.1007/BF02017220. Acesso em: 19 maio 2023.

OH, J.; LEE, B. G. A. Technical Approach for Suggesting Research Directions in Telecommunications Policy. **KSII Transactions on Internet and Information Systems**, [*S.l.*], v. 8, n. 12, p. 4467-4488, 2014. Disponível em: https://www.itiis.org/digital-library/manuscript/906. Acesso em: 19 maio 2023.

OKUBO, Y. **Bibliometric Indicators and analysis of research systems**: methods and examples. Paris: OECD Publishing, 1997.

OLIVEIRA, E. F. T.; 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://periodicos.ufmg.br/index.php/pci/article/view/22742. Acesso em: 19 maio 2023.

PINHEIRO, L. V. R.; FERREZ, H. D. **Tesauro Brasileiro de Ciência da Informação**. Rio de Janeiro; Brasília: Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), 2014.

PINTO, D. M. *et al*. O. Cartografia temática da produção técnico-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://rbbd.febab.org.br/rbbd/article/view/974/673. Acesso em: 19 maio 2023.

RODRIGUES, G. S.; AZEVEDO, R. A.; BATALHA, O. S. Produção científica em Arquivologia: uma abordagem quantitativa do Congresso Nacional de Arquivologia. **Revista Bibliomar**, São Luís, v. 20, n. 1, p. 31-56, jan./jun. 2021. Disponível em: http://periodicoseletronicos.ufma.br/index.php/bibliomar/article/view/16869. Acesso em: 19 maio 2023.

SANTOS, C. A. C. M. Organização e representação do conhecimento: contribuições aos estudos métricos. 2015a, **Anais...** Marília: FUNDEPE, 2015a. Disponível em: https://www.eca.usp.br/acervo/producao-academica/002790740.pdf. Acesso em: 19 maio 2023.

SANTOS, C. A. C. M. 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, 2015b. Disponível em: https://rbbd.febab.org.br/rbbd/article/view/494. Acesso em: 19 maio 2023.

SILVA, S. R. B.; CORREA, R. F. Sistemas de Indexação automática por atribuição: uma análise comparativa. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 25, p. 1-25, 2020. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2020.e70740. Acesso em: 19 maio 2023.

SILVA, S. R. B.; CORREA, R. F.; GIL-LEIVA, I. Avaliação direta e conjunta de Sistemas de Indexação Automática por Atribuição. **Informação & Sociedade:** estudos, João Pessoa, v. 30, n. 4, p. 1-27, 2020. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/57259. Acesso em: 19 maio 2023.

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: https://repositorio.ufmg.br/handle/1843/RRSA-6GGGUF. Acesso em: 19 maio 2023.

SOUZA, R. R. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 11, n. 1, p. 42-59, 2006. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2006v11nesp1p42. Acesso em: 19 maio 2023.

VAN ECK, N. J.; WALTMAN, L. Software survey: VOSviewer, a computer program for bibliometric mapping. **Scientometrics**, [*S.l.*], v. 84, n. 2, p. 523-538, 2010. Disponível em: https://link.springer.com/article/10.1007/s11192-009-0146-3. Acesso em: 19 maio 2023.

VAN ECK, N. J. *et al*. Automatic term identification for bibliometric mapping. **Scientometrics**, [*S.l.*], v. 82, n. 3, p. 581-596, 2010a. Disponível em: https://link.springer.com/article/10.1007/s11192-010-0173-0. Acesso em: 19 maio 2023.

VAN ECK, N. J. *et al*. A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. **Journal of the American Society for Information Science and Technology**, [*S.l.*], v. 61, n. 12, p. 2405-2416, 2010b. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/asi.21421. Acesso em: 19 maio 2023.

VERGARA, S. C. **Projetos e relatórios de pesquisa em administração**. 8. ed. São Paulo: Atlas, 2007.

NOTAS

AGRADECIMENTOS
O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

CONTRIBUIÇÃO DE AUTORIA
Concepção e elaboração do manuscrito: M. H. W. Ferreira, R. F. Correa
Coleta de dados: M. H. W. Ferreira
Análise de dados: M. H. W. Ferreira, R. F. Correa
Discussão dos resultados: M. H. W. Ferreira, R. F. Correa
Revisão e aprovação: M. H. W. Ferreira, R. F. Correa

CONJUNTO DE DADOS DE PESQUISA
Todo o conjunto de dados que dá suporte aos resultados deste estudo foi publicado no próprio artigo.

FINANCIAMENTO
Não se aplica.

CONSENTIMENTO DE USO DE IMAGEM
Não se aplica.

APROVAÇÃO DE COMITÊ DE ÉTICA EM PESQUISA
Não se aplica.

CONFLITO DE INTERESSES
Não se aplica.

PUBLISHER
Universidade Federal de Santa Catarina. Programa de Pós-graduação em Ciência da Informação. Publicação no Portal de Periódicos UFSC. As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES
Edgar Bisset Alvarez, Ana Clara Cândido, Patrícia Neubert, Genilson Geraldo, Mayara Medeira Trevilsom, Jônatas Edison da Silva, Camila Letícia Melo Furtado e Beatriz Tarré Alonso.

HISTÓRICO
Recebido em: 06-12-2022 - Aprovado em: 22-05-2023 - Publicado em: 28-06-2023.