UNIVERSIDADE FEDERAL DO TOCANTINS

CAMPUS UNIVERSITÁRIO DE PALMAS

PROGRAMA DE PÓS – GRADUAÇÃO DE

MESTRADO EM MODELAGEM COMPUTACIONAL DE SISTEMAS

RAFAEL MANSILHA MURTA

Crossed H-Index: uma ferramenta para investigar a autopromoção acadêmica em periódicos na base do Google Acadêmico

Palmas, TO
2023

Rafael Mansilha Murta

Crossed H-Index: uma ferramenta para investigar a autopromoção acadêmica em periódicos na base do Google Acadêmico.

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional de Sistemas da Universidade Federal do Tocantins, como requisito à obtenção do grau de Mestre em Modelagem Computacional de Sistemas.

Orientador: Dr. Waldecy Rodrigues

Palmas, TO
2023

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Tocantins

M984c MURTA, Rafael.

Crossed H-Index: uma ferramenta para investigar a autopromoção acadêmica em periódicos na base do Google Acadêmico . / Rafael MURTA. – Palmas, TO, 2023.

77 f.

Dissertação (Mestrado Acadêmico) - Universidade Federal do Tocantins – Câmpus Universitário de Palmas - Curso de Pós-Graduação (Mestrado) em Modelagem Computacional de Sistemas, 2023.

Orientador: Waldecy Rodrigues

1. Autocitação. 2. Bibliométrica. 3. Fator de Impacto. 4. H-index. I. Título

CDD 004

Elaborado pelo sistema de geração automática de ficha catalográfica da UFT com os dados fornecidos pelo(a) autor(a).

Rafael Mansilha Murta

Crossed H-Index: uma ferramenta para investigar a autopromoção acadêmica em periódicos na base do Google Acadêmico

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional de Sistemas da Universidade Federal do Tocantins, foi avaliado para a obtenção do título de Mestre em Modelagem Computacional de Sistemas e aprovado em sua forma final pelo Orientador e pela Banca Examinadora.

Data da aprovação: 01 / 03 / 2023 .

Banca Examinadora:

_____

Doutor Waldecy Rodrigues– UFT (Orientador)

_____

Doutor David Nadler Prata – UFT (Examinador Interno)

_____

Doutor Sérgio Oswaldo de Carvalho Avellar (Examinador Externo)

Dedico esse trabalho ao meu esposo Bruno, que nesses

10 anos juntos, sempre esteve ao meu lado,

te amo!

AGRADECIMENTOS

Agradeço à minha família, meu esposo Bruno, meus pais Rosângela e Claudio, minhas irmãs Jéssica e Janayna que sempre me incentivaram e permitiram que eu conseguisse concluir essa oportunidade. Ao orientador professor Dr. Waldecy Rodrigues e ao professor Dr. David Nadler Prata pelo excelente direcionamento e compartilhamento de conhecimento. Aos meus amigos, pela ajuda e incentivo. Minha psicóloga que me ajudou nesses últimos anos. Minha gratidão a todos que tiveram muita paciência comigo!

RESUMO

Esta pesquisa apresenta uma versão refinada da bibliométrica H-Index do Google Scholar Metrics (GSM) sem as autocitações, o Crossed H-Index. O processo de extração das autocitações dos autores consiste em comparar os nomes dos autores dos artigos que compõem o H-Index com os nomes dos autores das publicações que as citam, e verifica se a citação de um artigo é do mesmo periódico em que o artigo é publicado. Uma solução tecnológica foi implementada de forma a democratizar a consulta ao Crossed H-Index, disponibilizada no site da Web: http://palmas.uft.edu.br/crossed-h-index. Este trabalho conta também com uma análise das autocitações de 522 periódicos da área do conhecimento de Ciências Sociais Aplicadas, da subárea de Planejamento Urbano e Regional/Demografia. Destes, 64 periódicos obtiveram uma redução do seu H-index, em relação ao Crossed H-Index, superior à soma da média (13,89 %) com o desvio padrão (11,48 %) em relação ao grupo avaliado. Estes 64 periódicos sinalizam um possível abuso das autocitações, considerando que o impacto das autocitações em seu H-Index foge ao padrão dos periódicos no mesmo grupo da área do conhecimento.

Palavras-chave: Autocitação. Bibliométrica. Fator de Impacto. H-index. Crossed H-Index.

ABSTRACT

This research presents a refined version of the bibliometric H-Index from Google Scholar Metrics (GSM) without self-citations, the Crossed H-Index. The process of extracting the author's self-citations consists of comparing the names of the authors of the articles that compose the H-Index with the names of the authors of the publications that cites them, and verifying whether the citation of an article is from the same journal in which the article is published. A technological solution was implemented in order to democratize the consultation of the Crossed H-Index, available on the website: http://palmas.uft.edu.br/crossed-h-index. This work also has an analysis of the self-citations of 522 journals in the area of knowledge of Applied Social Sciences, in the subarea of Urban and Regional Planning/Demography. Of these, 64 journals obtained a reduction in their H-index, in relation to the Crossed H-Index, greater than the sum of the mean (13.89%) with the standard deviation (11.48%) in relation to the evaluated group. These 64 journals indicate a possible abuse of self-citations, considering that the impact of self-citations on their H-Index is beyond the standard of journals in the same group of knowledge area.

Key-words: Self-citation. Bibliometric. Impact Factor. H-index. Crossed H-Index.

LISTA DE ILUSTRAÇÃO

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPQ	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CSS	Folha de Estilo em Cascatas
CSIC	Conselho Superior de Investigações Científicas
IES	Instituição de Ensino Superior
FI	Fator de Impacto
GSM	*Google Scholar Metrics*
HTML	Linguagem de Marcação de HiperTexto
IES	Instituição de Ensino Superior
INPI	Instituto Nacional da Propriedade Industrial
ISIS	*Institute for Scientific Information*
ISSN	*Internacional Standard Serial Number*
JCI	*Journal Citation Indicator*
JCR	*Journal Citation Report*
MVC	*Model-View-Controller*
ORM	*Object Relational Model*
PHP	Hypertext Preprocessor
SIR	*SCImago Institution Rankings*
SJCR	SCImago Journal & Country Rank
SJR	*SCImago Journal Rank*
UFT	Universidade Federal do Tocantins
URL	Localizador Uniforme de Recursos
WoS	*Web of Science*

SUMÁRIO

1. INTRODUÇÃO

A ciência pode ser considerada como um amplo sistema social no qual uma de suas funções é disseminar o conhecimento. E uma das formas de se transmitir conhecimentos é por meio da divulgação científica: um trabalho de pesquisa deve ser publicado para que seus resultados sejam conhecidos no meio científico. Um dos mecanismos mais utilizados pela comunidade científica para a disseminação dos resultados das pesquisas é a publicação de artigos científicos em revistas, os chamados periódicos científicos e, para avaliar a produção científica de um determinado grupo de pesquisa foram elaborados indicadores para medir a sua visibilidade científica. Uma das ferramentas de estudo da cientometria são os índices bibliométricos, obtidos através de uma prática multidisciplinar, que começou a ser usada para identificar comportamento da literatura e sua evolução em contexto e época determinados que denomina-se bibliometria. Portanto, a bibliometria representa todos os estudos que tentam quantificar os processos de comunicação escrita fornecendo subsídios na formulação da política científica e tecnológica nas diferentes áreas do conhecimento (PIZZANI; SILVA; HAYASHI, 2008).

Na última década o mundo da pesquisa científica e tecnológica tem passado por transformações profundas, que têm exigido a adoção de novos instrumentos de intervenção e, em decorrência, o tratamento mais criterioso e coordenado da informação. Atualmente, a compreensão e análise dessa nova realidade, de sua dinâmica e complexidade, demanda a produção de indicadores mais robustos, que permitam, de um lado, a apreensão e interpretação de novas formas de produção, difusão e transferência de conhecimentos científicos e, de outro lado, a caracterização de maneira detalhada das capacidades nacionais em Ciência e Tecnologia - C&T no atual cenário mundial de desenvolvimento científico e tecnológico (SOARES et al., 2016).

Uma forma dos autores melhorarem seus conhecimentos é verificar uma pesquisa produzida em seu domínio de conhecimento: avaliando artigos submetidos a revista ou congresso científico, participando de bancas para avaliação de dissertação e mestrado ou teses de doutorados (SLOMSKI et al, 2013). Para Oliveira et al. (2013), o uso da pesquisa bibliométrica é um recurso precípuo para transmissão da produção científica e a sua finalidade é alcançada mediante a aplicação de uma técnica capaz de medir a influência dos pesquisadores ou periódicos, permitindo traçar o perfil e suas tendências, além de evidenciar áreas temáticas. A palavra da bibliometria é oriunda da fusão do sufixo “metria” e de bibliografia, informação,

ciência e biblioteca, sendo respectivamente análogos ou próximos de sua natureza, objetivos e aplicações (SANTOS, 2015).

Análises bibliométricas possibilitam estabelecer indicadores científicos de qualidade e confiabilidade, capazes de influenciar os processos envolvidos na recuperação e tratamento de dados e informações (GUEDES, 2012).

Para Araújo, Murakami e Prado (2018), o levantamento das citações de artigos para avaliação de impactos não permite uma visão ampla do real impacto dentro da comunidade acadêmica, pois, ao dar ênfase aos dados estatísticos, negligência os aspectos qualitativos da pesquisa, não sendo possível determinar se a produção científica em questão é verdadeiramente relevante academicamente.

Fischer et al. (2020a) explicam que a autocitação dentro da pesquisa científica, pode ser vista como instrumento de manutenção de um capital social em que alguns grupos fortalecem suas relações internas, mantendo relações de continuidade de junção de trabalhos e de percursos, que possam resultar em reconhecimento e inter-reconhecimento coletivo. Portanto, a autocitação é um fenômeno recorrente na produção acadêmica, não só na forma individual, mas também de forma colaborativa, apresentando conexões com outros fatores que permeiam a escrita acadêmica, como funcionamento nacional e internacional, com recursos, de grandes agências de fomento. Por essas razões, a autocitação merece ser estudada a fim de aprofundar compreensões em torno de práticas de escrita acadêmica, de respectivas publicações em periódicos e de índices de impacto dos textos em distintas áreas do conhecimento (FISCHER et al., 2020a).

A plataforma da SciELO admite trabalhos, onde para serem admitidos serão consideradas as citações nos índices bibliográficos, GSM e demais índices que porventura o periódico esteja indexado. Para os periódicos já indexados, o SciELO *Citation Index* que compreende os periódicos SciELO e *Web of Science* (WoS) é utilizado como índice de referência (SCIELO, 2017).

O fator de impacto indica o número médio de citações que os artigos de alguns periódicos recebem em um determinado ano, este índice serve para balizar a qualidade das publicações científicas e escolha de periódicos para os quais os autores querem submeter seus trabalhos (BARATA, 2010).

1.1 Justificativa

As alterações nas políticas de avaliação, da principal instituição de fomento à pesquisa no Brasil, com adesão ao uso de um índice bibliométrico que permite o uso da autocitação como uma forma de manipulação indevida do H-index, bibliométrica criada com o intuito de metrificar o impacto de uma produção científica.

1.2 Problema

Considerando que o processo analítico do GSM para a construção do seu H-index é completamente robotizado, o que consequentemente acarreta um índice propenso a falhas, e exploração de brechas no qual indivíduos podem tentar obter proveito.

1.3 Objetivos

1.3.1 Objetivo Geral

Esta pesquisa busca construir uma forma para avaliar as possíveis formas de autocitações presentes no H-index do GSM, de forma de refiná-lo e oferecer um índice mais refinado, no sentido do impacto em que periódicos possuem na comunidade acadêmica.

1.3.2 Objetivos Específicos

- A construção de um software que permita a consulta do Crossed H-index a partir da plataforma do GSM.
- Democratizar a consulta ao Crossed H-index, através da disponibilização da solução tecnológica implementada.
- Disponibilizar a ferramenta de forma prática e usual.
- Elaborar e disponibilizar a documentação da solução tecnológica.

1.4 Estrutura da dissertação

Esta dissertação está estruturada em 7 capítulos. O Capítulo 1, Introdução, apresenta uma contextualização do tema proposto deste trabalho, além de apresentar a Justificativa, definição dos objetivos.

O Capítulo 2 apresenta a Fundamentação teórica para com as bibliometria, principal tema abordado nesta dissertação, e suas definições. O Capítulo 3 apresenta a metodologia para

construção do Crossed H-Index e o Capítulo 4 descreve a solução tecnológica implementada ao problema apresentado.

No Capítulo 5 é apresentada uma análise realizada sobre a autocitação na área do conhecimento de Ciências Sociais Aplicadas, da subárea de Planejamento Urbano e Regional/Demografia. No Capítulo 6 são apresentadas as conclusões, seguido das referências e do Capítulo 7 com o Apêndice.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Bibliométricas

A bibliometria foi descrita originalmente como “bibliografia estatística”, termo utilizado por E. Wyndham Hulme (1923), sendo que a nomenclatura contemporânea por Paul Otlet (1934) foi cunhada em seu “Traité de Documentacion” (ARAÚJO, 2006). De acordo com Marcelo e Hayashi (2013) a bibliometria surgiu no início do século XX em resposta a necessidade de estudos e avaliações da produção científica, tendo como principal característica a elaboração de índices de produção do conhecimento científico.

Su e Lee (2010) explicam que a bibliometria tem sido usada como um método de análise quantitativa para pesquisa científica, logo os seus dados estatísticos elaborados por meio dos estudos bibliométricos medem a contribuição do conhecimento científico derivado das publicações em determinadas áreas, desta forma estes dados são usados na representação das atuais tendências das pesquisas e na identificação de temas para novas pesquisas. Para Francisco (2011) e Souza e Ribeiro (2013) a análise bibliométrica é uma técnica que tem como foco mensurar, de forma quantitativa, as publicações científicas de um autor ou Instituição de Ensino Superior (IES) em periódicos acadêmicos com seleção arbitrada por meio de padrões e métodos matemáticos e estatísticos.

Marcelo e Hayashi (2013) completam que apesar da bibliometria ter sua maior aplicação nos campos da Ciência da Informação, é possível aplicá-la em várias áreas do conhecimento com objetivo de avaliar o impacto da produção de uma determinada área de conhecimento, a produção e produtividade de um conjunto de investigadores, através da construção de indicadores bibliométricos. A ampla utilização de métodos e técnicas bibliométricas em diversas áreas do conhecimento deve-se aos avanços tecnológicos que atribuem uma maior apropriação dos fundamentos da bibliometria pelos pesquisadores, despertando o interesse de forma com que as pesquisas vêm crescendo amplamente a cada ano (MEDEIROS; VITORIANO, 2015).

A consolidação da Bibliometria no Brasil é evidenciada pelo crescente número de pesquisas que utilizam recursos e técnicas para obtenção de resultados que revelem indicadores de produtividade científica. Assim, a consolidação é firmada pela presença do tema em eventos nacionais da ciência da informação e de eventos específicos, que agregam pesquisadores da área (MEDEIROS; VITORIANO, 2015).

Stephan et al. (2017) explicam que os especialistas que integram comitês científicos de instituições de pesquisa em diversos países utilizam largamente indicadores bibliométricos baseados em citações, tais como: o Fator de Impacto (FI), H-index, e citações aferidas pelo GSM. Estes são utilizados como proxies para avaliar qualidade e impacto da pesquisa de candidatos a contratação e projeção na carreira.

Os índices bibliométricos proveem uma forma simples (em muitos casos, simplista) e conveniente de avaliar um grande número de candidatos, propostas ou artigos (NASSI-CALÒ, 2017). A autora completa que o FI e os indicadores similares de desempenho de periódicos possui suas limitações na função de avaliar artigos individuais e pesquisadores são conhecidas por todos. Os próprios pesquisadores contribuem para o cenário, causando um círculo vicioso, pois ao serem solicitados a identificar suas publicações mais relevantes, geralmente é selecionado com base no índice de citações, ao invés de selecionar artigos pela sua verdadeira importância acadêmica e científica ou uma pesquisa inovadora.

Conforme Guedes e Borschiver (2005), a bibliometria se fundamenta em três leis, que são consideradas basilares de sua estrutura: Lei de Bradford (produtividade de periódicos), Lei de Lotka (produtividade científica de autores) e Lei de Zipf (frequência de palavras).

Nesse contexto, a Bibliometria consiste em três leis fundamentais: Lei de Bradford, Lei de Lotka e Lei de Zipf. A lei de Bradford está relacionada com a dispersão da literatura periódica científica, sugerindo que a produtividade de artigos de determinado assunto poderá ser dividida em núcleos ou zonas, com intuito de medir a relevância de periódicos, em dada área do conhecimento (GUEDES; BORSCHIVER, 2005).

A bibliometria não é uma orientação homogênea, pois é dividida em pelo menos duas áreas: a descritiva, que trata aspectos quantitativos e a avaliativa, que acrescenta à primeira os estudos de avaliação da atividade científica (ÍÑIGUEZ-RUEDA et al., 2008). O levantamento do impacto acadêmico das publicações busca compreender o reconhecimento dos pesquisadores pela comunidade científica com base nas citações que os mesmos receberam (GONTIJO; ARAÚJO, 2021).

2.2 Autocitação

2.2.1 Definições de autocitação através das principais bases de periódicos

Segundo Spinak (2013) é considerado autoplágio, quando o autor plagia a si mesmo, reutilizando o seu próprio material, que já fora publicado outrora, sem indicar a referência de seu trabalho anterior. Desta forma, com maior rigor editorial, dependendo do contexto da publicação e da extensão do texto copiado, a inclusão da referência poderia ser suficiente, pois não indica claramente ao leitor e nem ao editor o alcance da cópia. Exposto isto, uma definição simples de autoplágio é usar a própria pesquisa anterior e apresentá-la para publicação como algo novo e original.

Fischer et al. (2020a) explicam que em relação ao fenômeno de autocitação foi possível observar que os artigos analisados mostraram produções colaborativas para a prática de autocitar trabalhos com parcerias, que se repetem nessas colaborações. Esse movimento demonstra um possível propósito, por parte dos autores, em tornar públicos trabalhos contínuos dentro do mesmo grupo de pesquisa, a fim de marcar confiabilidade frente ao leitor.

Além dos elementos apresentados até o momento em torno da autocitação, relacionados a ela estão, também, outros aspectos relevantes, como o fator de impacto e o H-index, os quais auxiliam na compreensão desta ocorrência em artigos de alto impacto (FISCHER et al., 2020b).

Apesar do exposto por Fischer et al. (2020a) em relação as produções colaborativas passarem uma imagem de autoridade sobre determinado assunto em relação a um grupo de pesquisadores, a Scielo (2017) explica que na avaliação do desempenho dos periódicos é observado o número de citações recebidas por artigo e o comitê consultivo analisa essa porcentagem de autocitação, é destacado que o elevado índice de autocitação em relação a outros periódicos de mesma área temática podem ser considerados prejudiciais na avaliação do periódico, podendo ser solicitado esclarecimentos do editor-chefe ou até mesmo a emissão de advertência e exclusão do periódico na coleção SciELO. Logo, o ideal é o periódico apresentar um índice de citações recebidas de artigos de outros periódicos e de autocitações compatível com os demais periódicos da mesma área temática (SCIELO, 2017).

2.2.2 JCR – *Journal Citation Reports*

A WoS é uma base de dados multidisciplinar que permite o acesso a 12.171 periódicos científicos, sendo editada pela empresa *Clarivate Analytics*. Isto é disponibilizado no Brasil por

meio do portal de periódicos do financiamento do governo brasileiro através da agência da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), a todas as instituições integrantes deste portal. O *Web of Science Journal Citation Report* (JCR) identifica e avalia os mais importantes periódicos de ciências e ciências sociais em todo o mundo e oferece análises sobre o desempenho dos periódicos (ATALLAH et al., 2020).

De acordo com Clarivate (2021), assim como outras métricas no JCR, o *Journal Citation Indicator* (JCI) é estático para o ano JCR. É criado a partir de um período dos dados, quando o JCR é produzido e inclui em seus dados as informações da Figura 1. O JCI é uma métrica normalizada por categoria, que é calculada para todos os periódicos do *WoS Core Collection*, incluindo o *Science Citation Index Expaned, Social Science Citation Index, Arts* & *Humanities Citation Index e Emerging Sources Citation Index.* Sendo esta outra métrica do impacto, que completamente o *Journal Impact Factor*, e que auxilia na avaliação de performance do periódico em um contexto definido.

Figura 1 – Produção do *Journal Citation Reports* (JCR).

Fonte: Clarivate (2021).

O JCR é um produto do *ISI Web of Knowledge*, sendo este um recurso autoritário para dados de fatores de impacto. Esse banco de dados fornece fatores de impacto e classificações de muitos periódicos nas ciências sociais e da vida com base em milhões de citações, este oferece várias opções de classificação, incluindo fator de impacto, total de citações, total de artigos e índice de imediatismo. Além disso, o JCR fornece um fator de impacto de cinco anos e dados de tendências visualizados (UC SAN DIEGO, 2021).

Conforme a UFMG (2022) o JCR fornece um modo sistemático e objetivo de avaliar os principais periódicos de pesquisa do mundo, pois oferece uma perspectiva exclusiva para

avaliação e comparação de periódicos por meio da cumulação e tabulação de contagens de citações e artigos de praticamente todas as especialidades nos campos da ciência, ciências sociais e tecnologia. Desta forma, o JCR apresenta: os periódicos mais frequentemente citados em um campo; os mais importantes em um campo; com maior impacto em um campo; os artigos mais publicados em um campo e dados da categoria do assunto para análise de desempenho.

2.2.3 SJR – *SCImago Journal Rank – Scopus*

O *SCImago Journal* & *Country Rank* (SJCR) é um portal disponível ao público que inclui os periódicos e indicadores científicos do país desenvolvidos a partir das informações contidas na base de dados Scopus® (Elsevier B.V.). O SJCR permite que você também incorpore métricas de diário significativas em sua web como um *widget* de imagem clicável. Esta plataforma leva o nome do indicador *SCImago Journal Rank* (SJR), desenvolvido pela SCImago a partir do algoritmo amplamente conhecido Google PageRank™. Este indicador mostra a visibilidade dos periódicos contidos na base Scopus® desde 1996. SCImago é um grupo de investigação do Conselho Superior de Investigações Científicas (CSIC), Universidade de Granada, Extremadura, Carlos III (Madri) e Alcalá de Henares, dedicado à análise, representação e recuperação de informação por meio de técnicas de visualização (SCIMAGO, 2022).

Além do Portal SJR, o SCImago desenvolveu o *The Shape of Science*, o *SCImago Institution Rankings* (SIR) e o *Atlas of Science*. *The Shape of Science* é um projeto de visualização de informação cujo objetivo é revelar a estrutura da ciência. Sua interface foi projetada para acessar o banco de dados de indicadores bibliométricos do portal *SCImago Journal Country Rank* resultados e impacto social medidos por sua visibilidade na web (SCIMAGO, 2022).

De acordo com Beatty (2016) o SJR pondera as citações com base na fonte de onde vêm, ou seja, o campo de assunto, a qualidade e a reputação da revista têm um efeito direto no valor de uma citação. O SJR também normaliza as diferenças no comportamento de citação entre os campos de assunto, além de calcular pelo SCImago Lab e ser desenvolvido a partir de dados do Scopus.

2.2.4 H-index

O uso da altmetria em complementaridade à bibliometria permite ampliar o escopo das análises do impacto acadêmico e social das produções científicas (GONTIJO; ARAÚJO, 2019).

O H-index foi desenvolvido por Jorge E. Hirsch no ano de 2005, da Universidade da Califórnia, este avalia a produção científica e mensura a relevância de publicações. Esse índice é um dos indicadores de destaque na literatura científica, que pode integrar o cálculo do fator de impacto, sendo este um parâmetro avaliativo robusto, pois somente considera aspectos relativos à produção de artigos e o impacto de número de citações de forma simultânea (SILVA; GRÁCIO, 2017).

Assim, segundo Vieira e Wainer (2013), ao se avaliar pesquisadores por indicadores bibliométricos como o H-index (combinação de produtividade e impacto), cada serviço como WoS, GSM ou *Scopus* possui suas próprias premissas e características de pesquisas.

Pois, indivíduos com índices H semelhantes são também comparáveis em termos de impacto científico, mesmo quando o número de artigos ou número total de citações de ambos for muito diverso. Diferentemente, quando comparado dois indivíduos (de idade científica igual), com números iguais de publicações ou citações e com índices H muito díspares, aquele com maior H-index é provavelmente um pesquisador mais talentoso (THOMÁS et al., 2011).

2.2.4.1 Google Acadêmico

O *Google Scholar* ou Google Acadêmico, como é conhecido no Brasil, busca referências em qualquer documento que esteja disponível na Web, inclusive versões eletrônicas de revistas e conferências, permitindo que qualquer documento citado por algum outro documento disponível na internet tenha suas citações monitoradas, e dispondo a partir deste monitoramento a funcionalidade de envio de notificações sobre assuntos do interesse do pesquisador (GOUVÊA et al., 2022).

Quanto ao Google Acadêmico, tem-se o *Google Scholar Metrics* (GSM), este fornece uma forma dos autores avaliarem de forma rápida a visibilidade e a influência de artigos mais novos em publicações acadêmicas. O *Scholar Metrics* pega as citações mais recentes de diversas publicações, para ajudar os autores a considerar onde publicar suas novas pesquisas. De forma sucinta, é possível navegar pelas 100 principais publicações em vários idiomas, ordenadas por suas métricas do H-index e mediana h dos últimos cinco anos. Com isto, é

possível verificar quais artigos de uma determinada publicação foram os mais citados e quem os citou (GOOGLE SCHOLAR, 2022).

Pérez-Rodríguez (2017) explica que a visibilidade da produção científica é um valor em ascensão na academia, logo a ferramenta GSM tornou-se um importante recurso tanto para pesquisadores quanto para os periódicos, pois permite estabelecer o índice de impacto de autores e publicações científicas, a partir das citações que são coletadas no GSM. Junto com WoS e Scopus, o *GSM* é referência para posicionar pesquisadores e periódicos em diferentes áreas do conhecimento, pois cada uma dessas bases de dados oferece indicadores e métricas que contribuem para a visibilidade dos periódicos. WoS e Scopus contam com os índices tradicionais mais prestigiados e conhecidos, com os quais têm sido avaliados os fatores de impacto dos periódicos através da recontagem de citações (JCR, SJR). Curiosamente, o GSM indexa mais que o dobro dos periódicos do que o SJR (Scopus) e dez vezes mais do que o JCR (WoS).

No caso do GSM, são analisados os periódicos e a produção científica dos autores a partir do índice bibliométrico h, calculado a partir do número de citações recebidas e da quantidade da produção científica de um pesquisador ou de um periódico, sempre aplicado a esse banco de dados nos últimos cinco anos. No entanto, o grande problema desta base de dados é que toda indexação é automática sem filtros e assim ocorre que são indexados em muitas ocasiões, documentos errados ou autores falsos e gera avaliações discutíveis. No entanto, essa ferramenta facilita a busca de autores conhecidos e sua classificação pelo h e o ranking de periódicos por idioma, ordenado pelo H-index (PÉREZ-RODRIGUEZ, 2017).

2.2.4.2 Clarivate

A *Clarivate Analytics*, empresa responsável pela base *Web of Science* e pela divulgação do FI, lançou no final de junho de 2019 o JCR 2019, com dados atualizados do FI das publicações. O JCR apresenta o FI de mais de 12 mil revistas indexadas na *Web of Science*. É publicado anualmente, com dados das publicações do ano anterior e citações dos dois anos imediatamente anteriores (UFRGS, 2020).

A *Clarivate Analytics* tem elaborado o perfil das principais universidades e instituições de pesquisa do mundo desde 2009, usando um conjunto exclusivo de indicadores-chave de desempenho para nossa iniciativa *Institutional Profiles*. A combinação de informações bibliométricas de padrão ouro com dados exclusivos sobre reputação, dados demográficos,

alunos e financiamento da equipe cria uma visão de 360 graus dos aspectos do desempenho de uma instituição. Perfis são utilizados por universidades, agências de financiamento, agências governamentais e rankings como uma ferramenta valiosa para identificar pontos fracos e fortes. O projeto *Institutional Profiles* (Perfis Institucionais) coleta, valida e analisa dados das três fontes exclusivas e de seis disciplinas acadêmicas (UNICAMP, 2022).

Embora o *Institute for Scientific Information* (ISIS) *Clarivate Analytics* seja o proprietário do FI e responsável pela sua publicação no JRC, as bases de dados Scopus e SciELO fornecem dados que possibilitam que os pesquisadores efetuem o seu cálculo (ALMEIDA; GRÁCIO, 2019).

2.2.4.3 Scopus H-index

A base de dados da Scopus é uma boa fonte de busca, pois apresenta organização da interface do sítio, sendo possível tornar a pesquisa mais didática e fluída, através de filtros dos resultados buscados, abertura e download de arquivos, formas de organização da listagem de arquivos de forma simples, clara e objetiva, evitando a possibilidade de haver retorno de artigos repetidos (FIGUEIREDO et al., 2017).

Scopus é uma base de dados disponível dentre as universidades, onde se obtém o H-index, é uma proposta para quantificar a produtividade e o impacto de pesquisas individuais ou em grupos baseando-se nos artigos (*papers*) mais citados. Por exemplo, um pesquisador com h=5 tem 5 artigos publicados que receberam 5 ou mais citações (UNESP, 2016).

A base de dados da Scopus inclui resumos e citações de literatura, abrangendo conteúdos nas seguintes áreas do conhecimento: química, física, matemática, engenharia, ciências da saúde e vida, ciências sociais, psicologia, economia, biologia, agricultura, ciências ambientais e ciências gerais (BAAS et al., 2021).

O H-index é um índice que tenta medir a produtividade e o impacto do trabalho publicado de um cientista ou acadêmico. No Scopus, o H-index não é um valor estático, este é calculado ao vivo em um conjunto de resultados cada vez que você a consulta. No Scopus você pode calculá-lo em qualquer conjunto de resultados; não precisa ser trabalhos pertencentes a apenas um autor. Basta executar uma pesquisa aleatória: *TITLE-ABS-KEY* (*mars water ice*), selecione todos os resultados, clique em *View Citation Overview* e será possível observar o valor do H-index para esse conjunto (ELSEVIER, 2014).

2.3 Impacto que bibliométricas tem em decisões de financiamento de pesquisa

De acordo com Oliveira (2018) a produção científica no Brasil é fruto da história das políticas científicas e do contexto sociocultural, uma vez que se concentra nas universidades públicas, o que a coloca no centro das políticas nacionais de ciência e tecnologia, logo é notável um crescimento expressivo da ciência brasileira nos últimos 25 anos.

Muitos autores defendem a publicação dos pareceres de forma aberta, alegando que esta prática pode ajudar a controlar fraudes no processo de avaliação, ao mesmo tempo em que permite o reconhecimento público dos pareceristas. Mais recentemente, pesquisadores têm discutido, qual seria o impacto de tornar o processo de *peer review* mais aberto e transparente. Ademais, a discussão foi ampliada para incluir a revisão de propostas de financiamento a pesquisa (*grants*). Este tema desperta a atenção tanto da comunidade científica quanto da sociedade civil científica (NASSI-CALÒ, 2015).

No processo de comunicação científica, é fundamental que as comunidades acadêmicas levantem e avaliem o desempenho de suas pesquisas por meio de indicadores apropriados, como os advindos dos estudos métricos da informação. A proposta das métricas alternativas em complementar as tradicionais para uma ampliação do âmbito de pesquisa, indo além dos dados de citações para análise de impactos e até mesmo ultrapassando o levantamento apenas no campo acadêmico, mostra-se importante para as comunidades científicas dos diversos campos do conhecimento (GONTIJO; ARAÚJO, 2021).

Para os pesquisadores, por garantir equidade e objetividade ao processo, e ao público em geral (contribuintes), para garantir que projetos sejam aprovados quanto ao mérito e não por motivos políticos ou para favorecer terceiros, aumentando a confiança da sociedade na pesquisa científica. A disponibilização dos pareceres pode abrir caminho para outro tipo de avaliação, por exemplo, aquela que ocorre após a publicação, a exemplo da revisão não anônima pós-publicação de artigos, como sugere Hilda Bastian, editora de *PLoS Medicine* e do *PubMed Commons*, uma iniciativa que permite postar comentários sobre trabalhos do PubMed após sua publicação. Em seu entender, comentários e críticas abertas podem melhorar a qualidade da pesquisa expondo erros e fraquezas de forma eficiente. No Brasil ainda prevalece o sistema cego (duplo ou simples) de revisão por pares em artigos de periódicos e propostas de financiamento à pesquisa, e a ideia de abrir os comentários dos revisores ainda não desperta interesse na comunidade científica (NASSI-CALÒ, 2015).

Define-se *grant* – auxílio à pesquisa – como um subsídio financeiro concedido por uma instituição governamental, privada ou sem fins lucrativos para o desenvolvimento de projetos. Não é esperado que os recursos recebidos sejam restituídos à fonte pagadora, que os utilizam para necessidades de infraestrutura, desenvolvimento de projetos de pesquisa, estudos, e outras atividades relacionadas. O sucesso na carreira de um pesquisador é geralmente medido por meio de publicações e métricas relacionadas ao seu impacto, como citações. É preciso considerar, no entanto, que sua habilidade em obter recursos para pesquisa é um fator preponderante para o progresso na carreira, pois está diretamente relacionada com sua capacidade de realizar as pesquisas e consequentemente, gerar publicações (NASSI-CAIÒ, 2019).

A avaliação de auxílio à pesquisa, por outro lado, tem influência mais direta em uma determinada área do conhecimento, por determinar se a pesquisa será financiada ou não e, neste último caso, poderá nunca se concretizar. Os critérios utilizados pelos pareceristas e o peso de cada um deles são, muitas vezes, decisões pessoais, ou podem ser formalmente providas pelos próprios financiadores. Por este motivo, estes pareceristas têm em suas mãos enormes desafios ao atribuir méritos a projetos, pois acabam tendo grande influência na pesquisa de toda uma área ou de um país (NASSI-CAIÒ, 2019).

O valor atribuído à pesquisa possibilita que as evidências produzidas contribuam para melhorar as políticas públicas (MORAES et al., 2019). Nos últimos anos, o interesse em demonstrar os benefícios das pesquisas e a necessidade de justificar o investimento público têm incentivado instituições e agências de fomento a desenvolverem sistemas de avaliação capazes de mostrar o impacto das pesquisas à sociedade (GREENHALGH et al., 2016). Mensurar o impacto desse investimento pode contribuir para aumentar a capacidade de financiar pesquisas estratégicas para o País, sobretudo em épocas de restrição orçamentária (MORAES et al., 2019).

Thomáz et al. (2011) explicam que os índices bibliométricos tem suas limitações, sendo assim, o ideal seria realizar o uso do conjunto de alguns destes índices para uma avaliação mais justa. Entretanto, apesar da subjetividade, a avaliação por pares ainda tem o seu valor, seja na avaliação de pesquisadores que se candidatam para cargos acadêmicos ou na avaliação editorial de artigos científicos. Logo, nenhum dos índices qualitativos e quantitativos, por melhor que sejam, é suficientemente preciso para ser utilizado de forma isolada, a combinação de alguns desses, associada à avaliação por pares, é certamente a melhor forma de avaliação objetiva.

2.4 Qualis / CAPES

O sistema Qualis foi introduzido em 1998 pela CAPES, uma agência do governo brasileiro cujo objetivo é avaliar e promover os programas de pós-graduação brasileiros. Desde o início, ficou claro que um dos principais aspectos de um programa de pós-graduação são as publicações de pesquisadores e estudantes. Uma vez que a quantidade de artigos produzidos é elevada, um sistema teve que ser desenvolvido para avaliar esta produção. Foi assim que o Qualis entrou no cenário (KELLNER, 2017).

A avaliação da qualidade dos periódicos, utilizados como meio de divulgação da pesquisa cientifica, vem sendo empregada como uma das formas de análise dos programas de pós-graduação, formando a conhecida lista *Qualis* da CAPES. O *Qualis* é o conjunto de procedimentos concebido para atender às necessidades específicas do sistema de avaliação, que cria e disponibiliza uma lista com a classificação dos veículos utilizados pelos programas de pós-graduação, para a divulgação da sua produção. Com o *Qualis* é possível verificar a qualidade dos artigos e de outros tipos de produção, a partir da análise da qualidade dos veículos de divulgação, ou seja, periódicos científicos e anais de eventos (THOMÁZ et al., 2011).

Este sistema classifica publicações em estratos que são utilizados para avaliar um determinado programa de pós-graduação (KELLNER, 2017).

2.4.1 Metodologia da avaliação quadrienal de 2017 a 2020

A metodologia utilizada pela CAPES para a última avaliação do Qualis, referente ao período de 2017 a 2020, foi baseada em cinco premissas com o objetivo de aprimorar o processo e proporcionar benefícios (XXXX, 2023). Sendo eles:

1. A classificação do periódico deve ser única, garantindo o mesmo status de qualificação da produção entre as áreas de avaliação;
2. Utilizar indicadores objetivos e desconsiderar fatores discricionários como pertinência ou relevância do periódico para a área;
3. O modelo não deve limitar o percentual de periódicos por estrato;
4. O modelo deve incorporar critérios de qualidade externos, ou seja, independente do uso que as áreas fazem dos periódicos;
5. O modelo deve ser indutor de internacionalização das publicações de artigos e também de indexação de periódicos.

O novo modelo de classificação de periódicos adota nove estratos, que são definidos com base nos percentis dos indicadores de cada publicação (XXXX, 2023):

- A1: percentil ≥ 87,5.
- A2: 75 ≤ percentil < 87,5.
- A3: 62,5 ≤ percentil < 75.
- A4: 50 ≤ percentil < 62,5.
- B1: 37,5 ≤ percentil < 50.
- B2: 25 ≤ percentil < 37,5.
- B3: 12,5 ≤ percentil < 25.
- B4: 0 ≤ percentil < 12,5.

Com base nos percentis dos indicadores de cada periódico, a metodologia de avaliação permite que as áreas mães façam ajustes nos estratos, limitados a 30%, a partir do estrato de referência. No entanto, os periódicos que não possuem indicadores ou não seguem as boas práticas editoriais definidas pela metodologia são enquadrados no estrato C e não receberam ajustes no estrato (XXXX, 2023).

3 METODOLOGIA

A fim de alcançar o objetivo proposto nesta pesquisa, um processo para extração do Crossed H-index do *GSM* foi elaborado. Este método consiste na construção de uma solução tecnológica para automatizar a coleta e estruturação das métricas e dos dados que compõem os índices H dos periódicos na base do *GSM*.

Após a coleta dos dados, por meio do *web scraping*, a solução tecnológica, calcula novamente o H-index dos periódicos, porém, sem utilizar as citações que são consideradas autocitações.

A sessão a seguir descreve cada etapa deste processo, assim como as especificações para construção protótipo da solução tecnológica, assim como a reformulação do protótipo inicial e a implementação de uma versão simplificada.

3.1 Coleta e estruturação dos dados

O processo de extração do Crossed H-index consiste em coletar as informações, relacionadas aos periódicos, disponíveis na página web de métricas do GSM [1]. Sendo elas: O nome do periódico, o valor do H-index, a mediana H, e quais as publicações que estão na lista que compõem o seu H-index.

Além dos dados do periódico, para cada publicação que compõem o H-index, é necessário coletar: O título da publicação, os seus autores, o periódico de publicação, o ano de publicação e o número de citações.

E, para cada artigo que compõem o H-index, é necessário coletar as informações referente as citações: O título da publicação que citou o artigo, os seus autores, o periódico em que foi publicado e o ano da publicação.

Este processo de estruturação dos dados a serem coletados para recalcular o H-index está representando na Figura 2 que é um modelo simplificado do diagrama de entidade relacionamento dos dados dos periódicos, publicações e citações.

[1] https://scholar.google.com.br/citations?view_op=metrics_intro&hl=pt-BR

Figura 2 - Modelo simplificado de entidade relacionamento do processo de coleta de informações das métricas dos periódicos no GSM.

Fonte: (Elaborada pelo autor, 2023).

3.1.1 Web Scraping

O primeiro passo neste processo de extração do Crossed H-index é a coleta das informações descritas acima. Todos estes dados estão disponíveis de forma pública na página do GSM, desta forma, a extração das informações será realizada por meio do uso do *web scraper*.

A necessidade de recorrer ao *web scraping*, se dá à ausência de uma interface de programação de aplicações, em inglês *Application Programming Interface* (API), para extração destas informações das métricas por parte do *GSM*.

O *web scraper*, ou coleta de dados web, consiste na mineração de dados disponíveis em sites da Internet. Uma das técnicas para o *web scraping* é o uso de um *web crawler*, também conhecido como extrator de dados da web. O *web crawler* é conceito de *software* que tem por objetivo extrair dados, ou informações, de uma página da Internet de forma automatizada, no qual o *software* atua como uma espécie de robô que acessa os links e coleta os dados. Estas informações podem ser utilizadas para as mais diversas finalidades.

O uso de um *software* automático para extração de informações, apesar de muito comum, pode ser visto como uma prática tortuosa, uma vez que dados podem ser coletados em

a ciência ou consentimento dos sites, contudo o próprio GSM já é uma plataforma em que todas as suas informações são coletadas por meio do *web scraping*.

Para (Fonseca, Magrani, & Campello, 2022), o *web scraping* não é uma prática expressamente definida com legal, ou ilegal, tendo em vista que não há uma regulamentação, ou legislação, específica sobre a prática. Mas, desde que conduzida de forma consciente e responsável, respeitando questões como: O direito à privacidade, à proteção dos dados, propriedade intelectual e termos de uso dos serviços, a prática é muito útil quando utilizada de forma consciente.

3.2 Recálculo do H-index

Após a coleta e estruturação das informações, o próximo passo é o recálculo do H-index. Nesta etapa o algoritmo da solução tecnológica percorre artigo por artigo de cada periódico e verifica, em cada citação, se há pelo menos um autor das citações possui o mesmo nome de algum autor da lista de autores da publicação citada, ou se o periódico, em que o artigo da citação foi publicado, possui o mesmo nome do periódico em que o índice está sendo recalculado.

A comparação de nomes dos autores e dos nomes dos periódicos, como critério de exclusão de autocitações, se dá ao fato que o critério de autocitação definido para esta pesquisa foi o de autores e periódicos. Esta escolha se dá devido ao fato de o Qualis, do sistema de avaliação da CAPES do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), ser a métrica mais popular no Brasil, a fim de avaliar a qualidade de produção acadêmica de um pesquisador brasileiro. Esta avaliação do pesquisador comumente é definida pelo Qualis dos periódicos em que o pesquisador possui publicações, desta forma, a principal finalidade desta métrica é qualificar os autores e periódicos.

Com o tratamento dos dados, após a remoção das autocitações detectadas de forma automática pelo *software* através da comparação de *strings*[2], o H-index é calculado com base na definição de Hirsch (HIRSCH, 2005) sem as autocitações de autores ou periódicos.

A página de métricas do GSM não disponibiliza todas as citações de todas as publicações dos periódicos, em sua listagem está disponível apenas até a linha de corte do H-

[2] *Strings*, ou cadeia de caracteres, é um tipo de dado formado pela sequência de caracteres. Este tipo de dado é muito utilizado para representar palavras, frases ou textos.

index, ou seja, se um periódico possui um H-index de 3, estão listadas as citações das 3 publicações mais citadas.

Esta limitação dificulta o processo de extração do crossed H-index, por exemplo, em um periódico com o H-index de 3, onde tanto a terceira quanto a quarta publicação mais citada possuem 3 citações, contudo, existe a possibilidade da terceira possuir autocitação e da quarta publicação mais citada, que também possui 3 citações, não conter autocitação. Desta forma, como o GSM não lista as publicações e citações após a linha de corte, existe a possibilidade do Crossed H-index não ser o real H-index sem autocitação do periódico.

Entretanto, como este processamento de dados visa apenas entender o cenário atual de autocitação, e não apresentar de fato qual seria o real H-index, além de que esta problemática pode causar uma variação de apenas nas publicações menos citadas, em sua maioria apresentando uma variação de apenas 1 no H-index, esta limitação não é considerada como prejudicial para a pesquisa.

4 SOLUÇÃO TECNOLÓGICA – CROSSED H-INDEX

A fim de atingir o objetivo proposto desta pesquisa, um software foi implementado a fim de facilitar o controle de autocitações dos periódicos. Este software foi elaborado no formato de *website*, de forma que não seja necessário instalar qualquer aplicativo e possa ser acessado de qualquer computador com internet.

Devido ao mecanismo de segurança dos navegadores modernos conhecido como *same-origin policy*, ao qual restringe como scripts de um site de uma origem podem interagir com os recursos de outro site com origem diferente, não é possível extrair o Crossed H-Index utilizando os próprios recursos do usuário, sendo necessário que a coleta das informações necessárias para construção do índice seja realizada por um servidor.

Para a construção deste servidor, a linguagem de programação *Hypertext Preprocessor*, ou PHP como é popularmente conhecida, foi escolhida devido ao PHP ter se consolidado por décadas disponibilização de sistemas para Web, o que alinha com o propósito da implementação. Outra razão é a familiaridade do desenvolvedor com a linguagem.

O *framework*[3] *Symfony* para PHP, ao qual utiliza o padrão de arquitetura de *software Model-View-Controller* (MVC), foi selecionado para implementação da solução tecnológica, no qual o *bundle*[4] *FriendsOfPHP*/*Goutte*[5], disponível para uso pela licença MIT[6], foi utilizado para realizar a ação de *web crawling*.

Utilizou-se o MariaDB, na sua versão 10.2.40, para o armazenamento dos dados necessários à extração do Crossed H-index. O seu gerenciamento é realizado pelo *framework* de mapeamento de objeto relacional, do inglês *object relational model* (ORM), Doctrine, por se tratar de uma ferramenta *Open Source* em PHP. A estrutura dos dados foi construída conforme a descrição das entidades na Figura 2.

O uso de um banco de dados para armazenamento destes dados, decorre da limitação de requisições, em um determinado período de tempo, que a plataforma de métricas do GSM

[3] *Framework*, dentro do contexto de desenvolvimento de sistemas, é um conjunto, ou estrutura, de códigos genéricos em que facilitam o processo de desenvolvimento.
[4] *Bundle* é um recurso disponível no *framework Symfony* semelhante aos *plugins*, permitindo que aplicações reaproveitem códigos/ferramentas da comunidade de desenvolvedores
[5] https://github.com/FriendsOfPHP/Goutte
[6] Criada pelo Instituto de Tecnologia de Massachusetts (MIT), é uma licença permissiva de *software*, podendo ser aplicada tanto em *software* livre, quanto em *software* proprietário.

impõem aos seus usuários. Desta forma o algoritmo realiza pausas entre cada requisição, por cerca de 10 segundos por cada Localizador Uniforme de Recursos (URL) acessado para coleta das informações.

Este processo pode demandar bastante tempo, uma vez que, quanto maior o H-index de um periódico, maior é o número de consultas, já que além da lista das publicações, é necessário acessar a URL de cada uma das listas de citações das publicações, e todas as listas são paginadas em 20 elementos.

Desta forma caso, para gerenciar a fila de requisições de Crossed H-Index, utilizou-se do servidor de código aberto de mensageria RabbitMQ, na sua versão 3.8, esta solução permite um controle da fila de trabalhos em *background*, possibilitando ao usuário desligar seu próprio equipamento sem prejudicar sua posição na fila de processamento.

Como o processo de coleta pode levar muito tempo, a persistência das informações foi a abordagem utilizada como uma forma de facilitar a consulta de periódicos que possuem muita procura pelo seu Crossed H-Index.

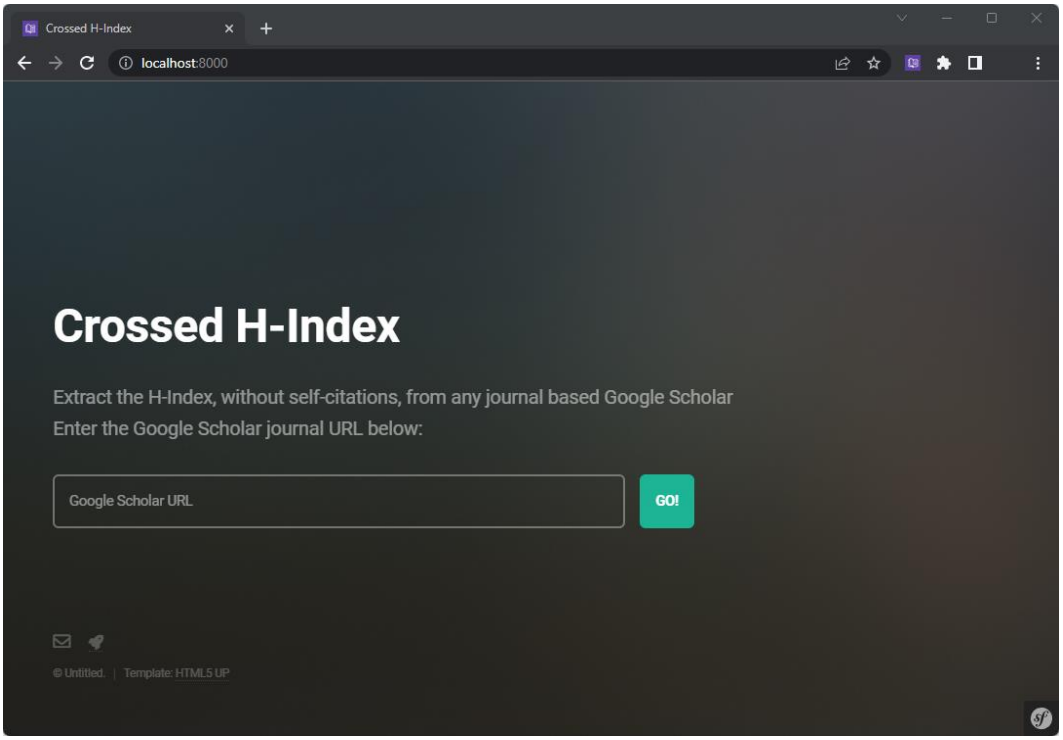
A solução tecnológica foi elaborada para os idiomas de Português Brasileiro e Inglês. Caso o navegador detecte que o usuário utiliza algum outro idioma que não seja os disponibilizados pela aplicação, por padrão, o site utilizará o idioma inglês.

A aplicação está registrada no Instituto Nacional da Propriedade Industrial (INPI) processo BR 51 2022 003022-5 apêndice 1.

4.1 Tela inicial da solução tecnológica

Visando atender ao objetivo de disponibilizar uma solução de forma prática e usual, a aplicação possui apenas duas telas, a primeira, representada pela Figura 3, possui apenas uma caixa de texto em destaque ao qual os usuários devem colocar uma URL da página de métricas de um periódico no GSM, a fim de consultar o seu Crossed H-Index.

Figura 3 - Tela inicial da solução tecnológica

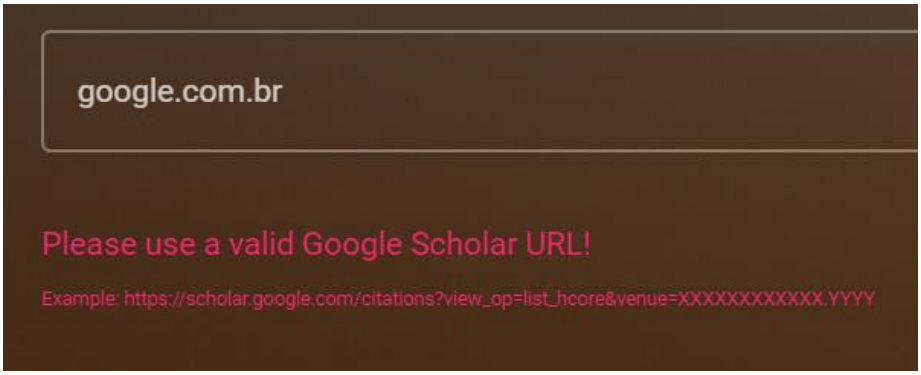


Fonte: (Elaborada pelo autor, 2023).

Nesta tela é realizada uma validação da URL informada pelo usuário em seu navegador, caso a URL não tenha o padrão necessário, conforme demonstrado na

Figura *4*, o sistema mostrará uma mensagem de erro, exemplificando como deve ser o formado da URL.

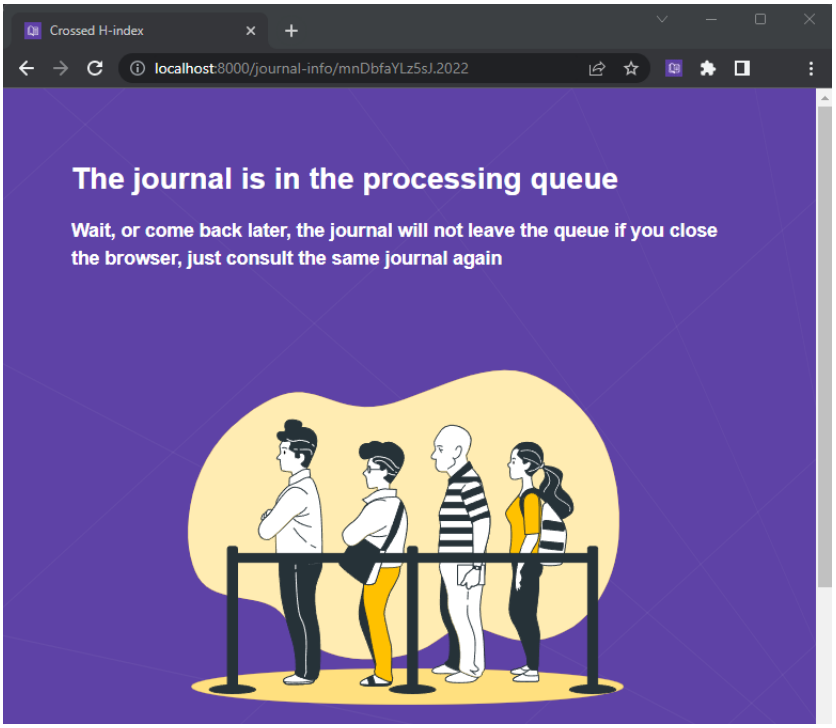
Figura 4 - Mensagem de URL errada



Fonte: (Elaborada pelo autor, 2023).

Caso a URL esteja de acordo com o formado necessário, a página solicita o Crossed H-Index do periódico e redireciona para a página de detalhamento, caso o Crossed H-Index do periódico desejado ainda não tenha sido extraído, sistema registrará na fila de processamento da solicitação de extração do Crossed H-Index e mostrará uma página, conforme representado na Figura 5, que será atualizada a cada minuto enquanto aguarda a extração do Crossed H-Index do periódico desejado.

Figura 5 - Página de espera da solicitação de extração do Crossed H-Index

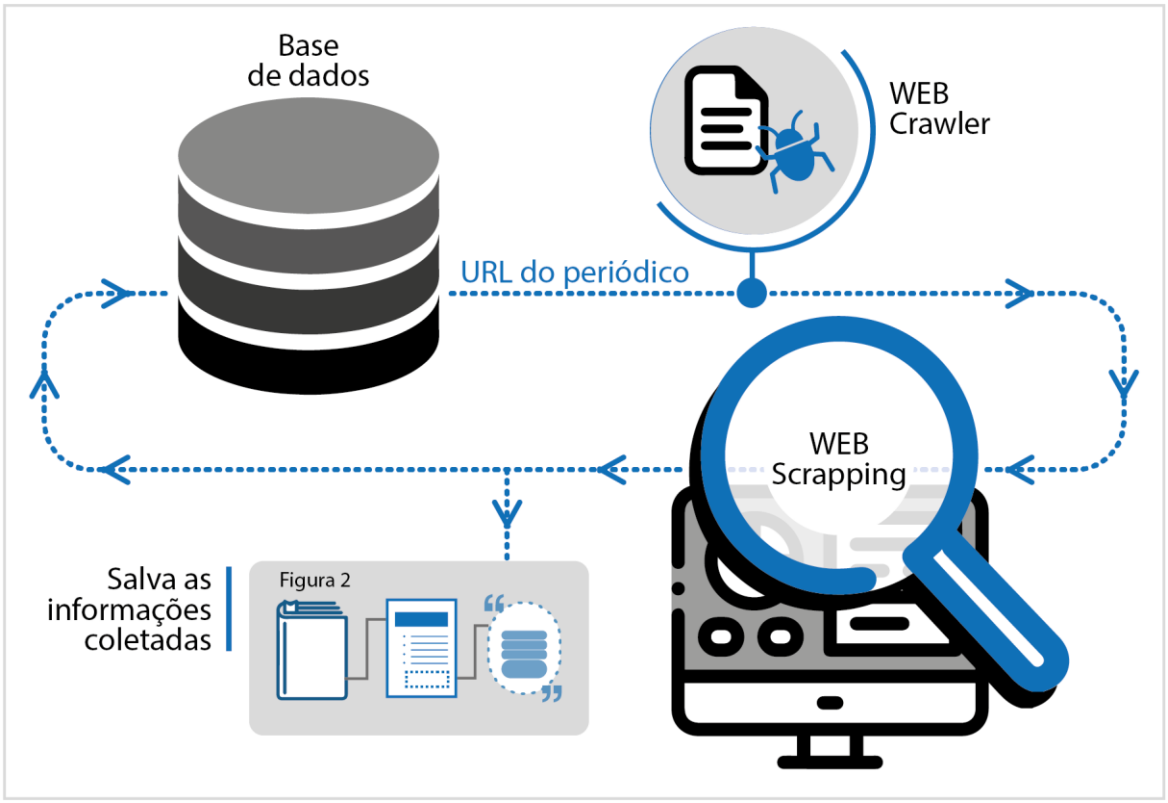


Fonte: (Elaborada pelo autor, 2023).

4.2 A extração do Crossed H-Index em *background.*

Por ordem de requisição, o servidor RabbitMQ executa o processo de extração do Crossed H-Index. Apenas uma requisição é processada por vez. O Script, conforme representado na Figura 6, usa um *web crawler* para percorrer todas as URLs dos periódicos e coletar todas as informações necessárias para extrair o Crossed H-index, como representadas na Figura 2.

Figura 6 - Representação do algoritmo para a coleta de dados necessários para extração do Crossed H-index.



Fonte: (Elaborada pelo autor, 2023).

Conforme representado na Figura 7, o *web crawler* coleta a página do periódico do qual são extraídas as informações:

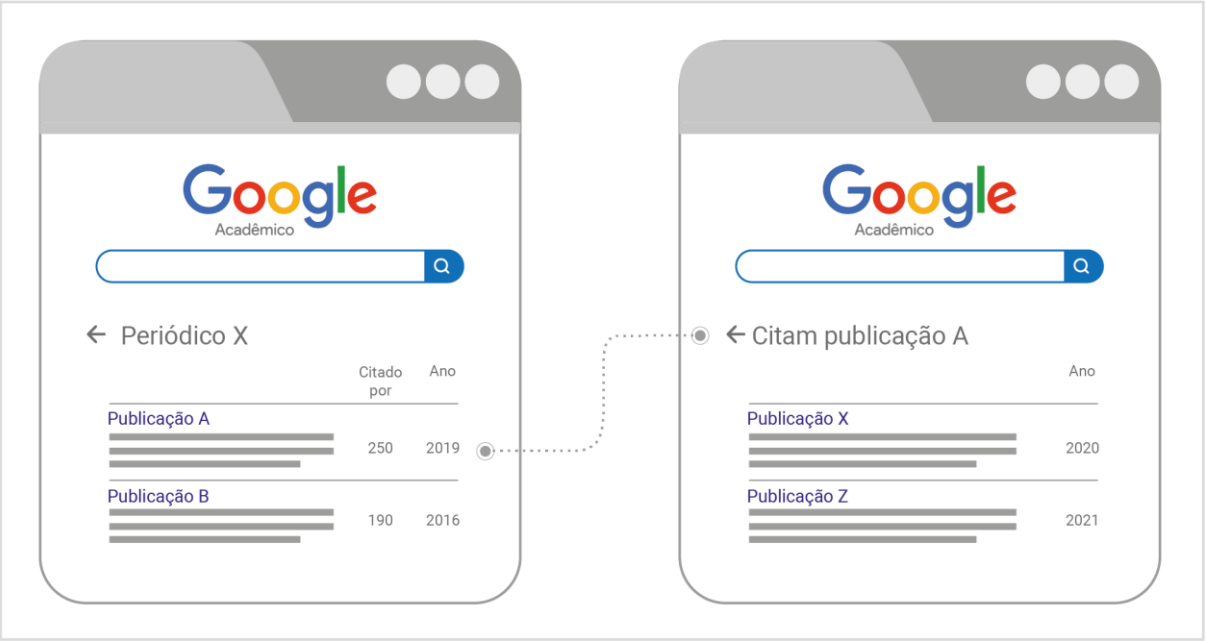
- Nome do periódico;
- H-index;
- Mediana H.

Para cada publicação o algoritmo coleta:

- Título da publicação;
- Nome do periódico da publicação (Para compor uma lista de possíveis nomes);
- Ano de publicação;
- Número de citações;

- Uma lista com o nome dos autores da publicação.

Figura 7 - Representação do processo de Web Scraping.



Fonte: (Elaborada pelo autor, 2023).

Mesmo que o nome do periódico já seja coletado uma vez, considerando que é possível que um periódico tenha alterado de nome, conforme detectado durante o processamento da análise preliminar, a solução tecnológica foi configurada para gerar uma lista com os possíveis nomes dos periódicos com base nos registros de onde o artigo por publicado que há em cada uma das publicações listadas quem compõem o H-index.

Para cada publicação, o sistema também coleta as seguintes informações em cada citação:

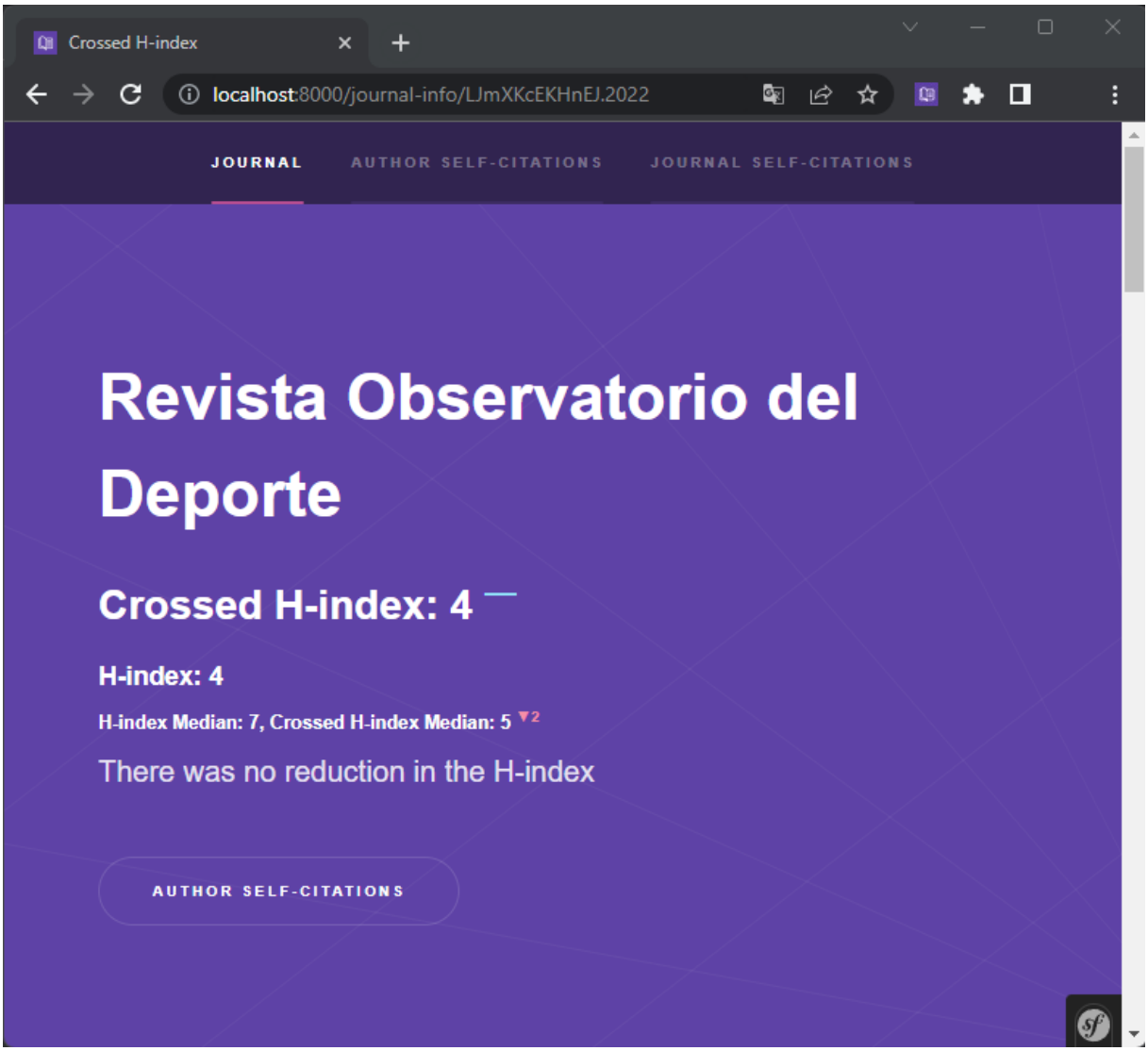
- Título da publicação que citou a publicação que compõem o H-index do periódico;
- Periódico em que citação foi realizada;
- Ano da citação;
- Lista com os nomes dos autores da publicação que citou.

4.3 Tela com detalhamento do Crossed H-Index do periódico.

A segunda tela da solução implementada é responsável pelo detalhamento do Crossed H-Index de um periódico. Caso a extração do Crossed H-Index ainda não tenha sido processada,

conforme representado pela Figura 5, apresentada uma mensagem solicitando ao usuário para aguardar enquanto a solicitação aguarda na fila de processamento.

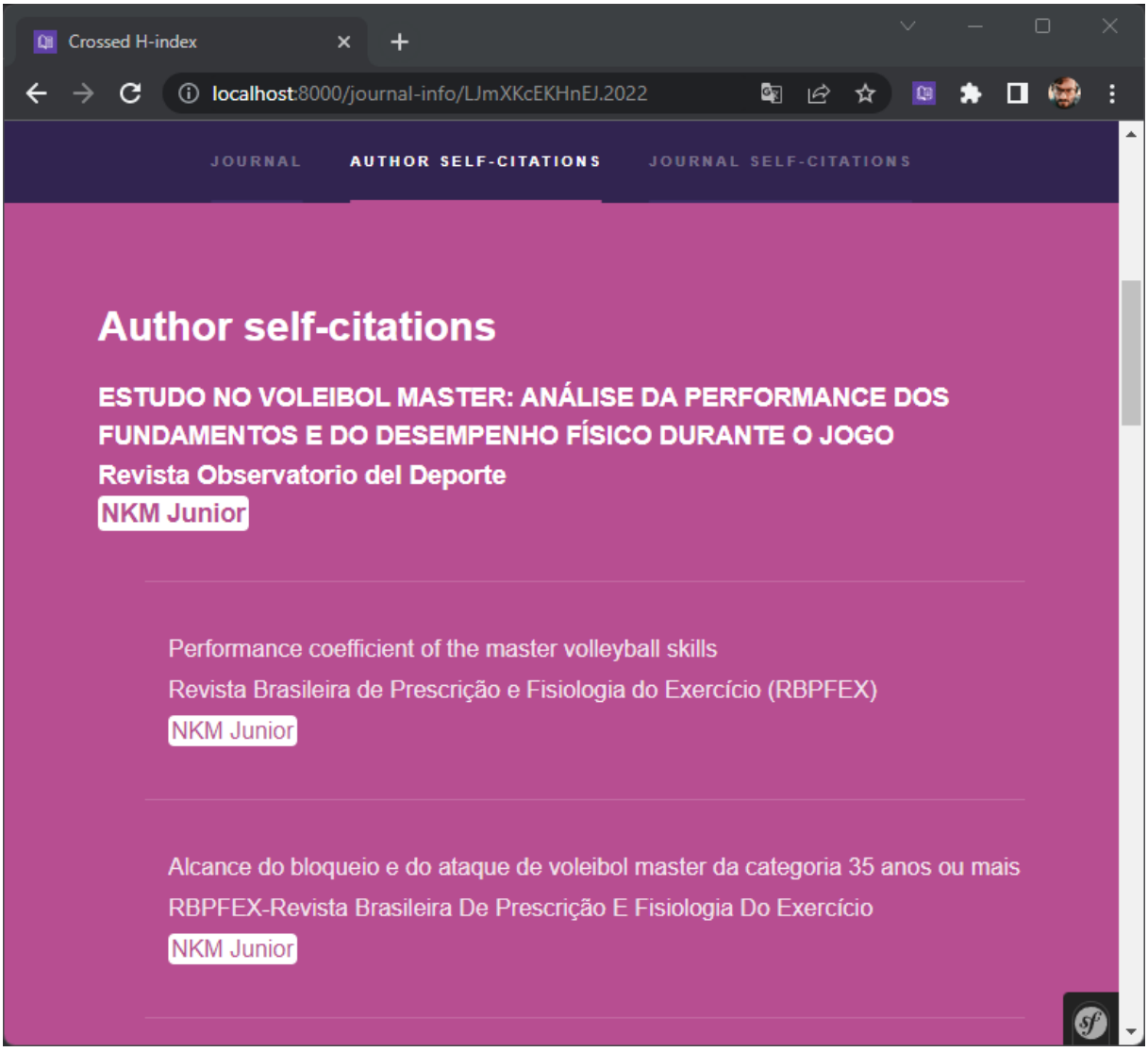
Figura 8 - Página do Crossed H-Index do periódico Revista Observatorio del Deporte



Fonte: (Elaborada pelo autor, 2023).

Caso a extração do Crossed H-Index já tenha sido processada, a página deve mostrar o detalhamento do Crossed H-Index do periódico, conforme apresentado na Figura 8, a página apresenta o valor em comparação ao H-Index original, nos casos em que há redução, este porcentual é apresentado.

Figura 9 - Detalhamento das autocitações



Fonte: (Elaborada pelo autor, 2023).

A página também apresenta as autocitações detectadas, dando destaque aos autores, ou ao periódico, em que houver identificação de autocitação. A Figura 9 é uma captura de tela da aplicação onde apresenta as autocitações identificadas para o autor NKM Junior. A escolha do periódico para demonstração das funcionalidades ocorreu de forma aleatória.

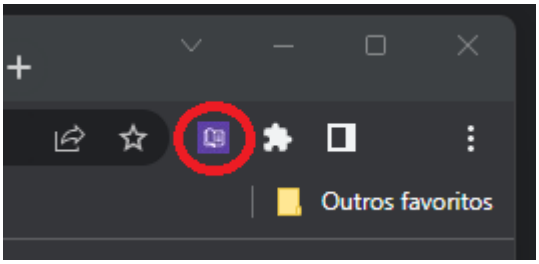
4.4 Extensão Crossed H-Index

Os navegadores modernos para computadores de mesa normalmente permitem o uso de extensões como uma forma de complementar a experiência do usuário ao navegar pelas páginas da web.

Diferentemente dos plug-ins que geralmente são executáveis, ou seja, arquivos binários, as extensões de navegadores são apenas código-fonte, portanto são auditáveis e permitem a identificação de códigos maliciosos com mais facilidade, desta forma os navegadores modernos tem removido o suporte à plug-ins.

Desta forma, a fim de facilitar o uso de analistas que trabalham com a avaliação de periódicos, uma extensão de navegador para a solução tecnológica foi implementado.

Figura 10 - Extensão Crossed H-Index instalada no navegador Google Chrome

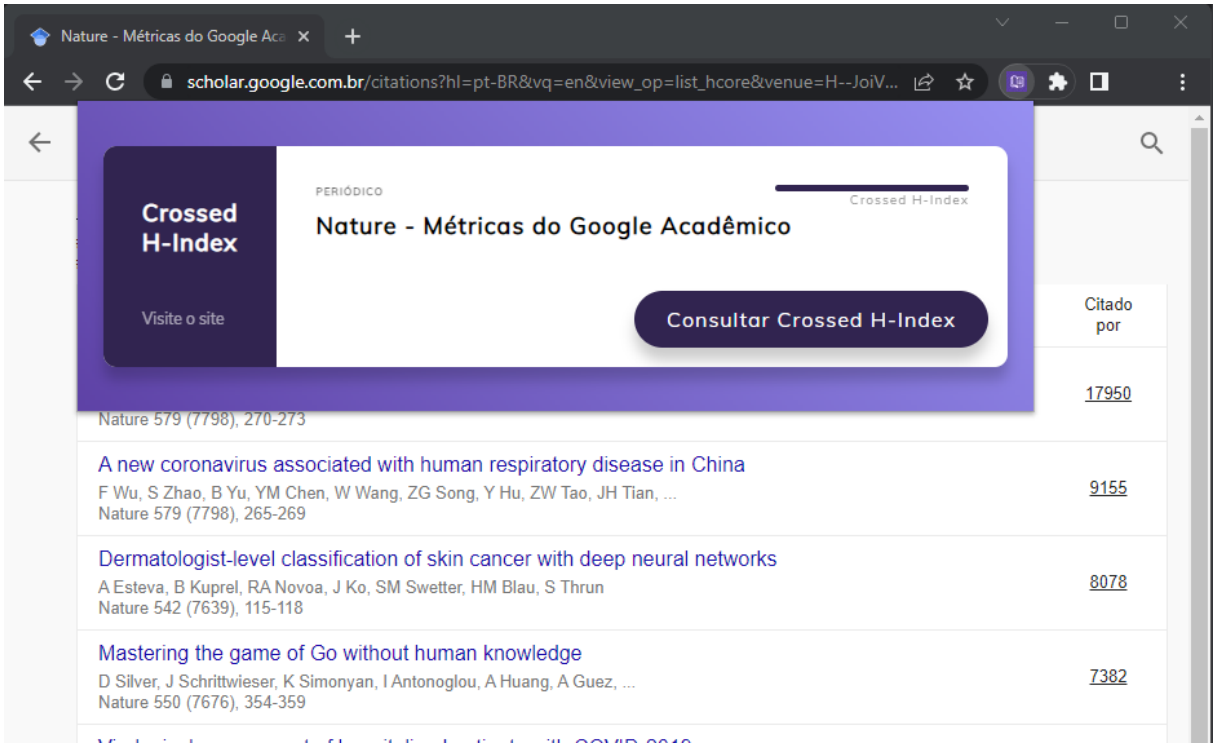


Fonte: (Elaborada pelo autor, 2023).

A extensão de navegador possui apenas duas telas, a primeira, representada pela Figura 11, aparece quando a extensão é aberta em uma página detalhada de métricas de algum periódico no Google Acadêmico.

Nesta tela a extensão habilita a opção de consulta ao Crossed H-Index do periódico. Ao clicar no botão, o navegador abrirá uma nova guia para a tela detalhada do Crossed H-Index do periódico, caso o Crossed H-Index ainda não tenha sido processado pelo servidor, o site adicionará o periódico à fila de extração do Crossed H-Index.

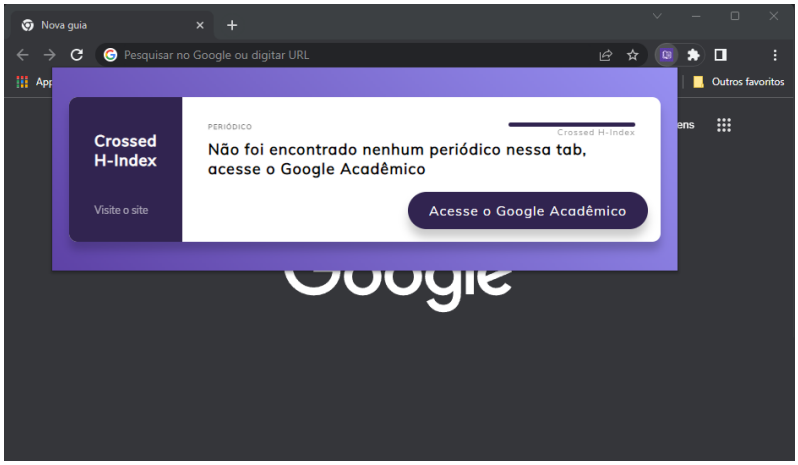
Figura 11 - Extensão Crossed H-Index na página de métricas do periódico Nature



Fonte: (Elaborada pelo autor, 2023).

A outra tela, representada pela Figura 12, aparece quando a extensão é aberta em qualquer página que não seja uma página de métricas de um periódico no Google Acadêmico, neste caso a extensão informa ao usuário que não foi encontrado nenhum periódico na guia e disponibiliza um link para acesso a página de métricas do Google Acadêmico.

Figura 12 - Extensão Crossed H-Index em uma página que não é a de métricas de um periódico no Google Acadêmico



Fonte: (Elaborada pelo autor, 2023).

A extensão de navegador foi elaborada para os idiomas de Português Brasileiro e Inglês. Caso o navegador detecte que o usuário utiliza algum outro idioma que não seja os disponibilizados pela extensão, por padrão a extensão utilizará o idioma inglês.

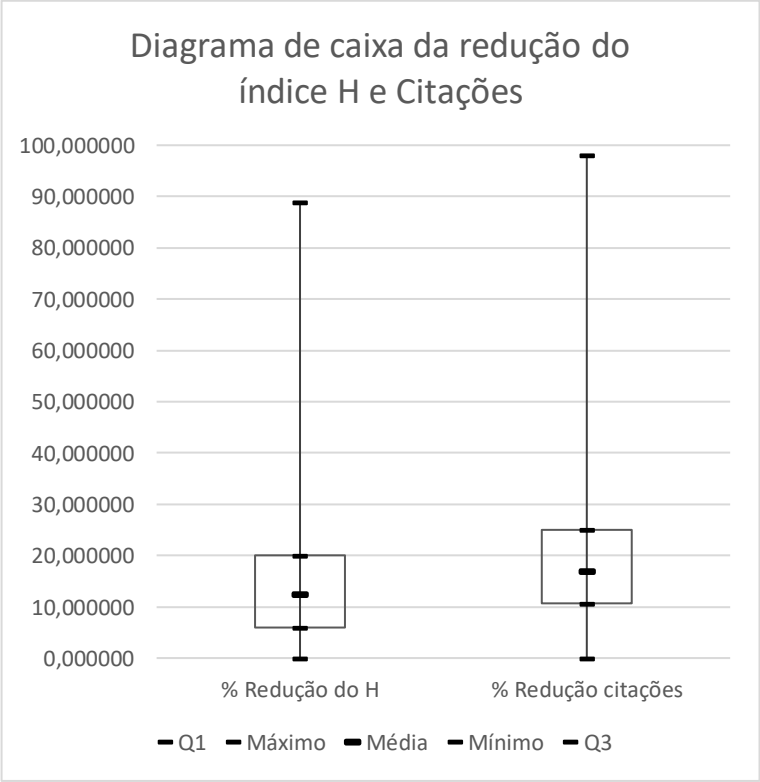
Para sua construção foram utilizadas apenas as linguagens Folha de Estilo em Cascatas (CSS), JavaScript e Linguagem de Marcação de HiperTexto (HTML). A aplicação está registrada no INPI BR 51 2022 003042-0 apêndice 2.

5 A AUTOCITAÇÃO NA ÁREA DE CONHECIMENTO DE CIÊNCIAS SOCIAIS APLICADAS, DA SUBÁREA DE PLANEJAMENTO URBANO E REGIONAL/DEMOGRAFIA

Visando validar o processo de extração do Crossed H-Index, para avaliar as autocitações, um total de 522 periódicos foram selecionados, como amostra, de uma população que compõem 1901 periódicos da área do conhecimento de Ciências Sociais Aplicadas, da subárea de Planejamento Urbano e Regional/Demografia. Considerando uma margem de erro de aproximadamente 5,00%, esta amostra apresenta um grau de confiança de 99,00%.

Após o recálculo do H-index, com base nas listagens de citações sem autocitação, dentre os 522 periódicos analisados o desvio padrão de redução do H-index dos periódicos foi de aproximadamente 11,5 % na redução do H-index em relação a seu valor original. O desvio padrão do percentual de redução do número total de citações dos periódicos foi de uma diminuição em aproximadamente 13,9 %.

Figura 13 - Diagrama de caixa da redução do H-index e Citações dos periódicos recalculados.



Fonte: (Elaborada pelo autor, 2023).

No diagrama de caixa, representado pela Figura 13, quanto a redução do H-index, a média geral de redução do H-index dos periódicos foi de 13,89 %. Ao segmentar em quartis, o Q1 ficou em aproximadamente 5,90 % e o Q3 em 20 %. Dentre os 522 periódicos, em que foram extraídos o Crossed H-Index, 292 estão entre o Q1 e o Q3.

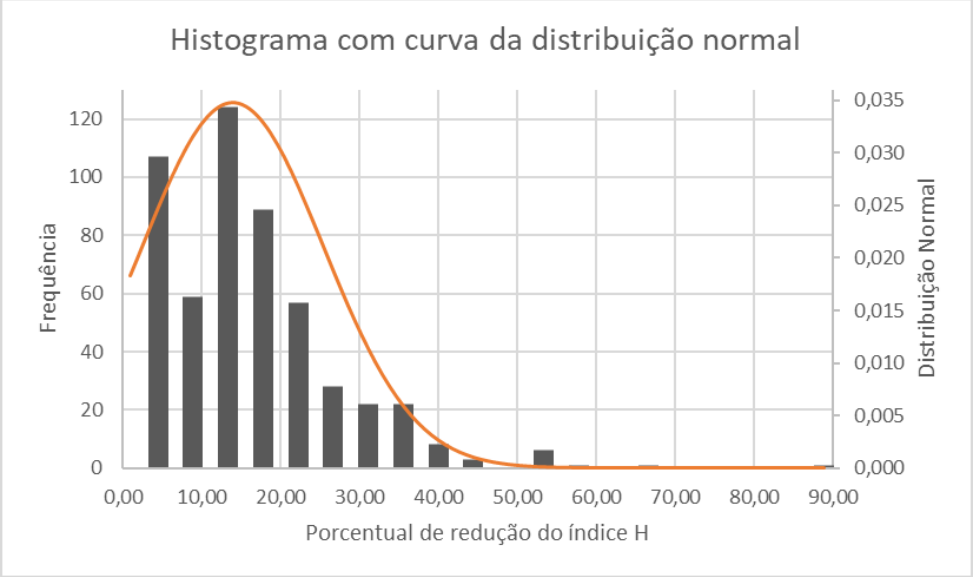
Entre o valor mínimo, de redução do H-index, e o Q1 estão 121 periódicos, dos quais, 99 periódicos obtiveram uma redução do H-index que corresponde com o mínimo de 0 %, pois, não houve redução dos seus H-index. Ou seja, são periódicos onde as autocitações não são suficientes para alterar o seu H-index, logo, periódicos neste grupo com um H-index maior possuem um grande impacto real na comunidade científica, uma vez que as citações dos trabalhos científicos só partem de outros pesquisadores e não dos próprios autores.

Outro aspecto observado na Figura 13, ainda relacionado ao H-index, é o tamanho da cauda entre o retângulo e os *outliers*. 109 periódicos estão na cauda entre o terceiro quartil e o valor máximo. O periódico com a maior redução seu H-index, que corresponde ao valor máximo de 88,89 %, teve uma redução do seu H-index em quase 4 vezes o valor de redução dos H-index de todos os periódicos abaixo do terceiro quartil, o que indica que há casos na ponta da cauda em que a prática de autocitação foge muito da normalidade.

No porcentual de redução do número de citações gerais, também representado pela Figura 13, a média de citações que foram identificadas como autocitações ficou em aproximadamente 17,07 % em relação ao número total de citações dos periódicos. Quando os dados são segmentados em quartis, o Q1 ficou em 10,67 % e o Q3 ficou em aproximadamente 25%.

Entre o valor mínimo, de citações detectadas como autocitação, e o Q1 estão 131 periódicos, dos quais, 32 periódicos não possuem nenhuma autocitação encontrada, desta forma o porcentual das autocitações corresponde ao valor mínimo de 0 %.

Figura 14 - Histograma do porcentual de redução do H-index com curva da distribuição normal.



Fonte: (Elaborada pelo autor, 2023).

Ainda relacionado as citações identificadas como autocitações para com a Figura 13, entre o Q1 e o Q3 estão 261 periódicos, desta forma, a cauda entre o retângulo e os *outliers* estão 130 periódicos em que foram encontrados casos de autocitações acima de 25 %.

A Figura 14 apresenta um histograma do porcentual de redução do H-index com a curva da distribuição normal. Assim como na Figura 13, a maior parte dos dados está concentrada próximo a média de 13,89%.

Nesta figura é possível reparar que o primeiro grupo, com uma linha de corte de redução do H-index de até 4,44 % contrariando com a representação gráfica da curva da distribuição normal, há um grupo de 106 periódicos, enquanto no segundo grupo com uma linha de corte acima de 4,44 % até 8,89 % estão 59 periódicos, seguindo pelo terceiro grupo de periódicos com uma redução do H-index acima de 13,33 % até 17,78%, grupo em que a média geral de redução do H-index de 13,89 %.

Ao observar o grupo de periódicos que obtiveram uma redução do H-index de até 4,44 %, representado pela Figura 14, constatou-se que este grupo possui uma média de citações antes do processo de extração de 384 citações sem autocitação por periódico. Após a filtragem das autocitações, a média de citações caiu para apenas 358, refletindo em uma média de autocitação nestes periódicos de 26 ocorrências.

Deste grupo de 106 periódicos, apenas 15 possuem até 10 citações, e 68 possuem até 100 citações antes do filtro de autocitações e 71 com até 100 citações após o filtro das autocitações. Desta forma, há um grupo de 35 periódicos com pelo menos 100 citações com uma média de autocitação muito baixa, quando comparado ao volume de periódicos no segundo grupo que está mais próximo à média geral.

5.1 Classificação dos periódicos de acordo com o grau de autocitação

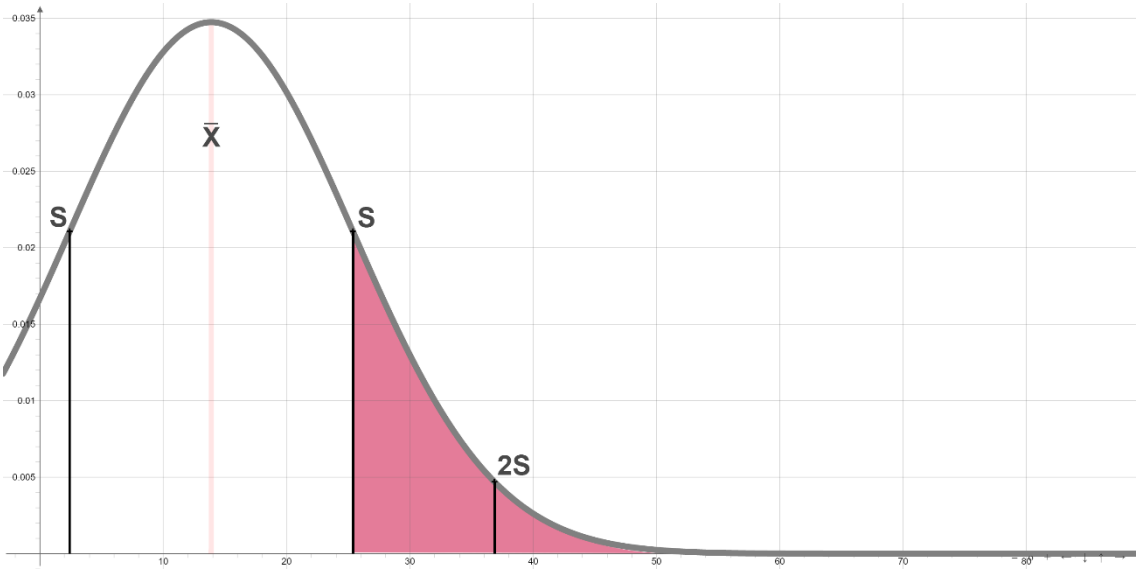
Dentre os 522 periódicos analisados, apenas em 32 não foram encontradas ocorrências de autocitação e consequentemente não tiveram seu H-index reduzido no processo de recálculo. Além destes, em 99 periódicos, apesar de 67 dos 99 apresentarem ocorrências de autocitação, os seus índices H não foram reduzidos após o recálculo.

Estes números enfatizam que a prática de autocitação não deve ser entendida necessariamente como uma atitude mal-intencionada, afinal, referenciar a trabalhos anteriores que se relacionam com o trabalho atual faz parte do processo científico, contudo, considerando que a citação é o principal fator para a grande maioria dos índices bibliométricos, abusos da prática de autocitação podem ser utilizados de forma mal-intencionada com o intuito de autopromoção acadêmica.

O critério utilizado como limite aceitável de autocitações é: Os periódicos cujo os porcentuais de redução dos seus índices H, após a remoção das autocitações, for menor que a soma da média amostral com o desvio padrão, e, portanto, são classificados no grupo azul.

Para os casos em quem o porcentual de redução é maior que a soma da média amostral com o desvio padrão (25,377 %), estes, são classificados no grupo vermelho. A Figura 15 é uma representação gráfica dos periódicos que estão classificados no grupo vermelho.

Figura 15 - Curva da distribuição normal com classificação dos periódicos.



Fonte: (Elaborada pelo autor, 2023).

Considerando o critério de não-conformidade, para os casos em quem o porcentual de redução é maior que a soma da média amostral com o desvio padrão, 64 periódicos foram classificados no grupo vermelho. Os índices H originais destes periódicos variavam entre 2 a 20, após a redução essa variação passou a ser de 1 a 14.

Na Tabela 1 a seguir, vemos uma lista dos 10 periódicos, classificados no grupo vermelho, com a maior redução do H-index, onde está variação do H-index foi igual, ou maior, a 4.

Tabela 1 - Lista dos dez periódicos, classificados no grupo vermelho com redução do H-index maior.

Identificação	Origem	Google H-index	Crossed H-index	Redução	Citações	Citações Refinadas	Autocitações	Porcentual de Autocitações
Periódico 1	Nacional	9	1	88,89 %	110	2	108	98,18 %
Periódico 2	Estrangeira	7	3	57,14 %	77	21	56	72,73 %
Periódico 3	Estrangeira	9	5	44,44 %	154	94	60	38,96 %
Periódico 4	Estrangeira	16	10	37,50 %	395	172	223	56,46 %
Periódico 5	Estrangeira	16	11	31,25 %	588	375	213	36,22 %
Periódico 6	Estrangeira	13	9	30,77 %	253	157	96	37,94 %
Periódico 7	Estrangeira	20	14	30,00 %	555	353	202	36,40 %
Periódico 8	Estrangeira	15	11	26,67 %	378	223	155	41,01 %
Periódico 9	Estrangeira	15	11	26,67 %	451	341	110	24,39 %
Periódico 10	Estrangeira	6	3	50 %	50	19	31	62 %

Fonte: (Elaborada pelo autor, 2023).

O periódico com o maior percentual de redução do H, de 9 para 1, com uma redução de aproximadamente 88,88%, teve 108 autocitações das 110 citações que concebiam o H-index original, ficando com apenas duas citações que não eram dos autores dos artigos, ou citações de publicações do próprio periódico. Na listagem da Tabela 1, este periódico é o único da que é de origem nacional.

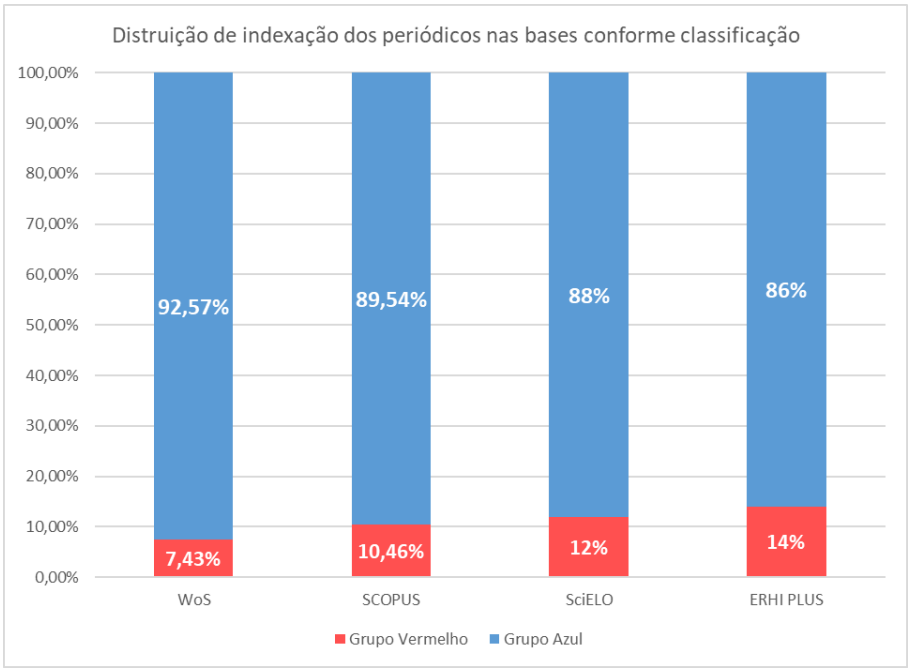
Tabela 2 . Bibliométricas dos dez periódicos classificados no grupo vermelho, com maior redução do H-index.

Identificação	ERHI PLUS	Google H-index	Crossed H-index	SJR H-index	Scopus CiteScore	Web of Science JCI
Periódico 1	-	9	1	-	-	-
Periódico 2	Indexado	7	3	-	-	-
Periódico 3	Indexado	9	5	19	1.4	0.14
Periódico 4	-	16	10	-	-	-
Periódico 5	-	16	11	18	3.4	0.79
Periódico 6	-	13	9	18	0.8	-
Periódico 7	-	20	14	18	1.8	-
Periódico 8	-	15	11	20	3.7	0.48
Periódico 9	Indexado	15	11	16	1.6	0.51
Periódico 10	-	6	3	4	1.1	-

Fonte: (Elaborada pelo autor, 2023).

A Tabela 2 apresenta as bibliométricas dos dez periódicos listados pela Tabela 1. Além da base do GSM, os periódicos foram consultados nas bases de dados: ERHI PLUS, Scopus, WoS e SciELO. A base de dados ERHI PLUS não possui bibliométrica, neste caso a coluna apenas aponta quais periódicos estão indexados na base. Os periódicos que não estão indexados nas bases não possuem nenhum valor nas bibliometrias.

Figura 9 – Distribuição de indexação dos periódicos nas bases conforme classificação



Fonte: (Elaborada pelo autor, 2023).

Dentre as bases descritas acima apenas a SciELO apresentou uma amostra muito pequena, onde apenas 25 periódicos dos 522 periódicos avaliados foram encontrados. Nenhum dos 9 periódicos listados nas Tabela 1 e Tabela 2 foram encontrados na base da SciELO.

A Figura 7 demonstra o porcentual de periódicos que estão indexados às bases conforme a classificação de conformidade e não conformidade com o número de autocitações. A base da WoS foi a base que obteve a maior diferença entre o número de indexações, onde, a cada 10 periódicos que possuem alto grau de autocitações, apenas 3 estão indexados à base, enquanto as bases Sopus e ERHI PLUS possuem cerca de 5 periódicos indexados às suas bases a cada 10 periódicos que possuem alto grau de autocitações.

6 CONCLUSÕES

Neste trabalho foi apresentado uma versão refinada da bibliométrica H-Index, o Crossed H-Index, que consiste no H-Index sem as autocitações de autor ou de periódico. O objetivo proposto de construir uma forma para avaliar as possíveis formas de autocitações presentes no H-index do GSM foi concluído através a construção de software que permita a consulta do Crossed H-index a partir da plataforma do GSM; Da democratização a consulta ao Crossed H-Index, através da disponibilização de um site da *Web*: http://palmas.uft.edu.br/crossed-h-index; A ferramenta foi construída de forma prática e usual e todo as funcionalidades e processos são descritos nesta dissertação. Desta forma os objetivos específicos descritos no Capítulo 1 foram concluídos.

As Bibliométricas são os principais parâmetros utilizado pelas instituições de ensino como critério para incentivo, principalmente financeiro, para o fomento de pesquisas. O H-Index do GSM, adotado em processos de avaliação de periódicos da principal instituição de fomento à pesquisa no Brasil, é construído em um processo completamente robotizado, o que permite o mal-uso da autocitação como uma forma de autopromoção acadêmica, o que foge ao propósito da bibliométrica de avaliar o impacto de uma determinada pessoa, ou periódico, na comunidade acadêmica.

É importante ressaltar que a autocitação não é necessariamente uma prática mal-intencionada. Ao longo da carreira de um pesquisador é muito provável que suas pesquisas estejam conexas, o que justifica o uso da autocitação, como uma forma de realizar essa conexão. Contudo esta autocitação em específico não representa um impacto dentro da comunidade científica, visto que não afetou outro indivíduo além do próprio autor.

A análise das autocitações na área de conhecimento de Ciências Sociais Aplicadas, da subárea de Planejamento Urbano e Regional/Demografia, através do parâmetro de avaliação da redução do H-Index superior à soma da média (13,89 %) com o desvio padrão (11,48 %) do periódico, para com o grupo de periódicos da mesma área do conhecimento, apresentou 63 de 522 com um índice considerado elevado de autocitações, levando em consideração o porcentual de redução do seu índice H.

Esta pesquisa indica que H-index do GSM é um bom medidor de qualidade dos periódicos da área do conhecimento de Ciências Sociais Aplicadas, da subárea de Planejamento Urbano e Regional/Demografia, contudo o processo totalmente automatizado de busca das

citações, permite o uso de autocitações como forma de inflar seu índice bibliométrico, uma vez que a plataforma não disponibiliza o H-index com um tratamento das autocitações.

Diferente da bibliométrica do H-index, determinada pela base GSM da Alphabet, os outros índices bibliométricos comumente utilizados pela comunidade acadêmica, determinados pelas bases, como, Scopus da Elsevier e Web Of Science da Clarivate, possuem uma análise mais subjetiva que vai além do número de citações, seus índices tendem a serem mais refinados que algoritmos que fazem uma contagem robótica do número de citações. Tendo em vista que o H-Index do GSM é utilizado para construção de índices qualitativos de produção científica, o Crossed H-Index apresenta-se como uma solução para esta questão, servido de complemento para os processos de avaliação.

REFERÊNCIAS

ALMEIDA, C. C. de.; GRÁCIO, M. C. C. Produção científica brasileira sobre o indicador "Fator de Impacto": um estudo nas bases SciELO, Scopus e Web of Science. **Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v.24, n.54, p.62-77, 2019.

ARAÚJO, Carlos Alberto. Bibliometria: evolução histórica e questões atuais. **Revista em Questão**, v. 12, n. 1, p. 11-32, 2006.

ARAÚJO, R. F. de.; MURAKAMI, T.; PRADO, J. A repercussão de artigos de periódicos brasileiros da ciência da informação no Facebook: um estudo altmétrico. **Revista Digital de Biblioteconomia e Ciência da Informação**, v. 16, n. 12, p. 365-379, 2018.

ATALLAH, A. N.; PUGA, M. E. S.; AMARAL, J. L. G. Web of Science Journal Citation Report 2020: the Brazilian contribution to the "Medicine, General & Internal" category of the journal impact factor (JIF) ranking (SCI 2019). **São Paulo Medical Journal**, n.138, v.4, p.271-274, 2020.

BAAS, J.; BOYACK, K.; IOANNIDIS, J. P. A. **August 2021 data-update for "Updated science-wide author databases of standardized citation indicators".** Elsevier BV, 2021. Disponível em: https://elsevier.digitalcommonsdata.com/datasets/btchxktzyw/3. Acesso em: 29 de jun. de 2022.

BARATA, G. F. **Nature e Science: mudança na comunicação da ciência e a contribuição da ciência brasileira (1936-2009).** 2010. 247 f. Tese (Doutorado em História) - Faculdade de Filosofia, Ciências e Letras, Universidade de São Paulo, São Paulo. 2010.

BEATTY, S. **Journal Metrics in Scopus: SCImago Journal Rank (SJR).** Elsevier, 2016. Disponível em: https://blog.scopus.com/posts/journal-metrics-in-scopus-scimago-journal-rank-sjr. Acesso em: 19 de jun de 2022.

CASS, S.; KULKARNI, P.; GUIZO, E. **Top Programming Languages - IEEE Spectrum.** IEEE Spectrum, 2022. Disponível em: https://spectrum.ieee.org/top-programming-languages/. Acesso em: 20 de jun. de 2022.

CLARIVATE. **Journal Citation Indicator**. 3 f. 2021. Disponível em: https://www.periodicos.capes.gov.br/images/documents/Journal%20Citation%20Reports%20JCR%20(guia).pdf. Acesso em: 29 de jun. de 2022.

ELSEVIER. **The Scopus h-index, what's it all about? Part I.** Elsevier, 2014. Disponível em: https://blog.scopus.com/posts/the-scopus-h-index-what-s-it-all-about-part-i#:~:text=The%20h%2Dindex%20is%20an,time%20you%20look%20it%20up.Acesso em: 20 de jun de 2022.

FIGUEIREDO, A. R.; WANDERLEY B. G.; VILAS BOAS, T. S.; SANTOS, M. C. Estudo da eficiência dos portais Science Direct, Scopus, Lilacs e Periódicos CAPES, evidenciando seus aspectos positivos e negativos. **Scientia Amazonia**, v.6, n.2, p.1-10, 2017.

FISCHER, A.; FERREIRA, K. M.; SILVA, R. da. Escrita acadêmica em artigos científicos: autocitação em diferentes áreas disciplinares. **Revista Online de Política e Gestão Educacional**, v.24, n.3, p.1257-1271, 2020a.

FISCHER, A.; GRIMES, C.; KOSLOSKI, E. R.; VICENTINI, M. A. Padrões da autocitação em artigos de alto impacto da revista nature. **RIAEE – Revista Ibero-Americana de Estudos em Educação**, v.16, n.1, p.276-291, 2020b.

FONSECA, C.; MAGRANI, E.; CAMPELLO, T. **Web scraping: legal ou ilegal?** 2022. Disponível em: https://politica.estadao.com.br/blogs/fausto-macedo/web-scraping-legal-ou-ilegal/. Acesso em: 20 de jun. de 2022.

FRANCISCO, E. de R. RAE-eletrônica: exploração do acervo à luz da bibliometria, geoanálise e redes sociais. **Revista de Administração de Empresas**, v. 51, n. 3, p. 280-306, 2011.

GONTIJO, M. C. A.; ARAÚJO, R. F. de. Impacto acadêmico e atenção on-line de pesquisas sobre inteligência artificial na área da saúde: análise de dados bibliométricos e altmétricos. **Revista Eletrônica de Biblioteconomia e Ciência da Informação**, v.26, e76249, 2021.

GONTIJO, M. C. A. Dados bibliométricos e altmétricos de artigos científicos sobre inteligência artificial, análise do impacto acadêmico e social. **Múltiplos Olhares em Ciência da Informação**, v. 9, n.2, p.1-11, 2019.

GOOGLE SCHOLAR. **Google Scholar Metrics.** 2022. Disponível em: https://scholar.google.com/intl/en/scholar/metrics.html. Acesso em: 21 de jun. de 2022.

GOUVÊA, A. L.; ÁVILA, C. H. de.; LADISLAU, D. O.; LIMA, G. M. de.; RIBEIRO, G. H. M.; VAZ, J. A.;TOMAZ, L. B. P.; SOARES, M. D.; NETTO, N. M.; COSTA, P. R.; SILVA, R. I. da.; VIEIRA, W. S. G.; GAYDÈCZKA, B.; OKURA, M. H.; MALPASS, A. C. G.; MALPASS, G. R. P. Índice H dos pesquisadores brasileiros: um olhar comparativo entre as bases de dados WoS, Scopus e Google Scholar. **Research, Society and Development,** v.11, n.5, e13711527832, 2022.

GREENHALGH, T.; RAFTERY, J.; HANNEY, S.; GLOVER, M. Research impact: a narrative review. **BMC Medicine**, v.14, n.1, p.1-16, 2016.

GUEDES, V. L. S. A bibliometria e a gestão da informação e do conhecimento científico e tecnológico: uma revisão da literatura. **Ponto de Acesso**, v. 6, n. 2, p.74-109, 2012.

GUEDES, V. L. S.; BORSCHIVER, S**. Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica.** Anais do Encontro Nacional de Ciências da Informação, v. 6, p.1-18, 2005. Disponível em:< http://www.cinform-anteriores.ufba.br/vi_anais/docs/VaniaLSGuedes.pdf>. Acesso em: 27. jun. 2022.

HIRSCH, J. E. An index to quantify an individual's scientific research output. **Physical Sciences**, v.102, n.46, p.16569-16572, 2005.

ÍÑIGUEZ-RUEDA, L.; MARTINEZ-MARTINEZ, L. M.; MUÑOZ-JUSTICIA, J. M.; PEÑARANDA-COLERA, M. C.; SAHÚN-PADILLA, M. A.; ALVARADO, J. G. The Mapping of Spanish Social Psychology through its conferences: a bibliometric perspective. **The Spanish Journal of Psychology**, v.11, n.1, p.137-158, 2008.

KELLNER, A. W. A. The qualis system: a perspective from a multidisciplinar journal. **Anais da Academia Brasileira de Ciências**, v.89, n.3, p.1339-1342, 2017.

MARCELO, J. F.; HAYASHI, M. C. P. I. Estudo bibliométrico sobre a produção científica no campo da sociologia da ciência. **Informação & Informação,** v.18, n.3, p.138-153, 2013.

MEDEIROS, J. M. G. de.; VITORIANO, M. A. V. A evolução da bibliometria e sua interdisciplinaridade na produção científica brasileira. **Revista Digital de Biblioteconomia e Ciência da Informação**, v. 13, n.3, p.491-503, 2015.

MORAES, L. H.; ÂNGULO-TUESTA, A.; FUNGHETTO, S. S.; REHEM, T. C. M. S. B. Impacto das pesquisas do Programa de Apoio ao Desenvolvimento Institucional do Sistema Único de Saúde. **Saúde Debate,** v.43, n.especial 2, p.63-74, 2019.

NASSI-CALÒ, L. **A miopia dos indicadores brasileiros.** SciELO em perspectiva, 2017. Disponível em: https://blog.scielo.org/blog/2017/06/01/a-miopia-dos-indicadores-bibliometricos/#.Yrs1MXbMKUl. Acesso em: 25 de jun. de 2022.

NASSI-CALÒ, L. **Os pareceres de propostas de financiamento a pesquisa poderiam ser abertos?** SciELO em perspectiva, 2015. Disponível em: https://blog.scielo.org/blog/2015/03/20/os-pareceres-de-propostas-de-financiamento-a-pesquisa-poderiam-ser-abertos/#.YrDt53bMLIU. Acesso em: 20 de jun de 2022.

NASSI-CALÒ, L. **Um olhar sobre a avaliação por pares de propostas de auxílio à pesquisa.** SciELO em Perspectiva, 2019. Disponível em: https://blog.scielo.org/blog/2019/12/04/um-olhar-sobre-a-avaliacao-por-pares-de-propostas-de-auxilio-a-pesquisa/#.YrDt-nbMLIU. Acesso em: 15 de jun. de 2022.

OLIVEIRA, E. F. R. de. **Estudos métricos da informação no Brasil: indicadores de produção, colaboração, impacto e visibilidade.** Oficina Universitária: São Paulo: Cultura Acadêmica, 184 p. 2018.

OLIVEIRA, S. C. M.; BARBOSA, E. S.; REZENDE, I. C. C.; SILVA, R. P. A.; ALBUQUERQUE, L. S. **Bibliometria em artigos de contabilidade aplicada ao setor público.** In: CONGRESSO BRASILEIRO DE CUSTOS – Uberlândia – MG, 2013.

PIZZANI, L.; SILVA, R. C. da.; HAYASHI, M. C. P. I. Bases de dados e bibliometria: a presença da educação especial na base Medline. **Revista Brasileira de Biblioteconomia e Documentação**, v.4, n.1, p. 68-85, 2008.

PÉREZ-RODRIGUEZ, M. A. Google Scholar Metrics (GSM). **Revista Comunicar**, 2017. Disponível em: https://www.revistacomunicar.com/wp/school-of-authors/google-scholar-metrics-gsm/. Acesso em: 21 de jun. de 2022.

SANTOS, G. C. dos. Análise Bibliométrica dos Artigos Publicados como Estudos Bibliométricos na História do Congresso Brasileiro de Custos. **Pensar Contábil**, v. 17, n. 62, p. 4-13, 2015.

SCIMAGO, (n.d.). **SJR — SCImago Journal & Country Rank [Portal]**. Retrieved, 2022. Disponível em: https://www.scimagojr.com/aboutus.php. Acesso em: 29 de jun. de 2022.

SLOMSKI, V. G.; PINHEIRO, I. C. B.; MEGLIORINI, E.; FERREIRA, T. A. R. C. **A importância da formulação da questão de pesquisa na produção científica em contabilidade: uma discussão a partir de trabalhos publicados no Congresso Brasileiro de Custos no ano de 2009.** In: CONGRESSO BRASILEIRO DE CUSTOS – Uberlândia - MG, 2013. Disponível em: https://anaiscbc.emnuvens.com.br/anais/article/view/117/117. Acesso em: 29 de jun. de 2022.

SCIELO – Scientific Electronic Library Online. **Critérios, política e procedimentos para a admissão e a permanência de periódicos científicos na coleção SciELO <país>.** SciELO, 30 f. 2017.

SILVA, D. D; GRÁCIO, M. C. C. Índice h de Hirsch: análise comparativa entre as bases de dados Scopus, Web of Science e Google Acadêmico. **Em Questão**, v. 23, p. 196-212, 2017.

SOARES, P. B.; CARNEIRO, T. C. J.; CALMON, J. L.; CASTRO, L. O.C. O. Análise bibliométrica da produção científica brasileira sobre Tecnologia de Construção e Edificações na base de dados Web of Science. **Ambiente Construído**, v. 16, n. 1, p. 175-185, 2016.

SOUZA, M. T. S. de; RIBEIRO, H. C. M. Sustentabilidade ambiental: uma meta-análise da produção brasileira em periódicos de administração. **Revista de Administração Contemporânea**, v. 17, n. 3, p. 368-396, 2013.

SPINAK, E. **Ética editorial e o problema do autoplágio [online].** SciELO em Perspectiva, 2013. Disponível em: https://blog.scielo.org/blog/2013/11/11/etica-editorial-e-o-problema-do-autoplagio/. Acesso em: 20 de jun. de 2022.

STEPHAN, P.; VEUGELERS, R.; WANG, J. Reviewers are blinkered by bibliometrics. **Nature,** v.544, n.7651, p.411-412, 2017.

SU, H.; LEE, P. Mapping Knowledge Structure by Keyword Co-Occurrence: a first look at journal papers in technology foresight. **Scientometrics**, v. 85, n. 1, p.65-79, 2010.

THOMÁZ, P. G.; ASSAD, R. S.; MOREIRA, L. F. P. Uso do fator de impacto e do índice H para avaliar pesquisadores e publicações. **Arquivos Brasileiros de Cardiologia**, v.96, n.2, p.90-93, 2011.

UC SAN DIEGO. **Measuring your Research Impact: Journal Citation Reports (JCR).** 2021. Disponível em: https://ucsd.libguides.com/ResearchImpact/JCR. Acesso em: 29 de jun. de 2022.

UFMG – Universidade Federal de Minas Gerais. **Journal Citation Reports (JCR).** 2022. Disponível em: http://biblioteca.qui.ufmg.br/?q=node/29. Acesso em: 29 de jun. de 2022.

UFRGS – Universidade Federal do Rio Grande do Sul. **Fator de Impacto: como verificar a métrica e o ranking de periódicos no JCR.** Bliblioteca Setorial ICBS, 2020. Disponível em: https://www.ufrgs.br/bibicbs/fatordeimpacto/. Acesso em: 20 de jun. de 2022.

UNESP – Universidade Estadual Paulista. **Passo a passo para obter o índice h – Scopus.** UNESP, 3f, 2016. Disponível em: https://www.fclar.unesp.br/Home/Biblioteca/indice-h-scopus.pdf. Acesso em: 27 de jun. de 2022.

UNICAMP – Universidade Estadual de Campinas. **Clarivate Analytics**. 2022. Disponível em: http://www.prdu.unicamp.br/rankings/clarivate-analytics. Acesso em: 27 de jun. de 2022.

VIEIRA, P. V. M.; WAINER, J. Correlações entre a contagem de citações de pesquisadores brasileiros, usando o Web of Science, Scopus e Scholar. **Perspectivas em Ciência da Informação**, v.18, n.3, p.45–60, 2013.

CAPES - Diretoria de Avaliação/CAPES. Documento técnico do Qualis Periódicos. 2023. Disponível em: https://www.gov.br/capes/pt-br/centrais-de-conteudo/documentos/avaliacao/avaliacao-quadrienal-2017/DocumentotcnicoQualisPeridicosfinal.pdf. Acesso em: 02 de março de 2023.

7 APÊNDICE

7.1 Registro de Software – Crossed H-Index

	Criador: GLEYTON DE CASTRO LIMA; JONAS LOTUFO BRANT DE CARVALHO; JORGE HENRIQUE CABRAL FERNANDES; MARCELA LOPES SANTOS; ROGÉRIO LUIZ ARAÚJO CARMINÉ Linguagem: JAVA; JAVA SCRIPT; POSTGREE SQL; PYTHON; SQL Campo de Aplicação: SD-01; SD-02 Tipo de Programa: GI-01 Data da Criação: Não informado
Processo: BR 51 2022 003021-7	Código 730 - Expedição do Certificado de Registro Título: Solving one-dimensional two-phase flow problems in rigid porous media using L-scheme Titular: MARCELO HENRIQUE SABATINI Criador: MARCELO HENRIQUE SABATINI; MARCIO AUGUSTO VILLELA PINTO; MICHELY LAÍS DE OLIVEIRA; SEBASTIÃO ROMERO FRANCO Linguagem: MATLAB Campo de Aplicação: FQ-05; FQ-06; MT-04; MT-05; MT-06 Tipo de Programa: SM-01 Data da Criação: 19/11/2020
Processo: BR 51 2022 003022-5	Código 730 - Expedição do Certificado de Registro Título: Crossed H-Index Titular: RAFAEL MANSILHA MURTA Criador: RAFAEL MANSILHA MURTA Linguagem: CSS; JAVA SCRIPT; PHP; SHELL SCRIPT Campo de Aplicação: IF-02 Tipo de Programa: AP-01; GI-04 Data da Criação: 26/10/2022
Processo: BR 51 2022 003023-3	Código 730 - Expedição do Certificado de Registro Título: HIGEST Titular: EMANUEL PINHEIRO PEQUENO Criador: EMANUEL PINHEIRO PEQUENO Linguagem: OUTROS Campo de Aplicação: SD-07; SD-08 Tipo de Programa: AP-01 Data da Criação: 11/05/2021
Processo: BR 51 2022 003024-1	Código 730 - Expedição do Certificado de Registro Título: HIGEST Titular: EMANUEL PINHEIRO PEQUENO Criador: EMANUEL PINHEIRO PEQUENO Linguagem: OUTROS Campo de Aplicação: SD-07; SD-08 Tipo de Programa: AP-01 Data da Criação: 11/05/2021
Processo: BR 51 2022 003025-0	Código 730 - Expedição do Certificado de Registro Título: SENTNEL Titular: SENTNEL MONITORAMENTO LTDA

7.2 Registro de Software – Extensão Crossed H-Index

	Linguagem: FORTRAN Campo de Aplicação: CC-05; IN-03 Tipo de Programa: SM-01 Data da Criação: 01/09/1995
Processo: BR 51 2022 003040-3	Código 730 - Expedição do Certificado de Registro Título: ALTERMOB - Plataforma de Mobilidade Urbana Titular: ALTERCAR TECNOLOGIA Criador: MOISES MOISÉS DE JESUS DA SILVA MIRANDA Linguagem: CSS; HTML; JAVA; JAVA SCRIPT; MYSQL; PHP Campo de Aplicação: AD-10; AD-11; ED-06; UB-04 Tipo de Programa: AP-01 Data da Criação: 10/07/2022
Processo: BR 51 2022 003041-1	Código 730 - Expedição do Certificado de Registro Título: FERRAMENTA PARA PROTECAO INSTANTÂNEA DE DADOS PESSOAIS EM CELULARES Titular: SIMBUS NEGOCIOS E CORRETORA DE SEGUROS LTDA Criador: MÁRCIO GABRIEL SOUZA VALIM Linguagem: JAVA SCRIPT Campo de Aplicação: IF-04; IF-09; IF-10 Tipo de Programa: AP-01; IA-02; SO-06 Data da Criação: 30/09/2022
Processo: BR 51 2022 003042-0	Código 730 - Expedição do Certificado de Registro Título: Extensão Crossed H-Index Titular: RAFAEL MANSILHA MURTA Criador: RAFAEL MANSILHA MURTA Linguagem: CSS; HTML; JAVA SCRIPT Campo de Aplicação: IF-02 Tipo de Programa: GI-04 Data da Criação: 26/10/2022
Processo: BR 51 2022 003043-8	Código 730 - Expedição do Certificado de Registro Título: SiSale SNGPC Titular: SÉRGIO BARROS DE SOUSA Criador: SÉRGIO BARROS DE SOUSA Linguagem: PHP Campo de Aplicação: SD-10 Tipo de Programa: AP-01; AT-03; GI-01; UT-01 Data da Criação: 29/10/2022
Processo: BR 51 2022 003045-4	Código 730 - Expedição do Certificado de Registro Título: SFGestorPublico Titular: SÓFOLHA SOLUÇÕES CORPORATIVAS LTDA Criador: WANDERLEY NOVELLO DE LIMA JUNIOR Linguagem: DELPHI Campo de Aplicação: AD-01; AD-02; AD-04; AD-05; AD-08; AD-09; AD-11; EC-04; FN-01

7.3 Artigo – Uma rede social construída a partir de documentos digitais do portal da Universidade Federal do Tocantins.

UMA REDE SOCIAL CONSTRUÍDA A PARTIR DE DOCUMENTOS DIGITAIS DO PORTAL DA UNIVERSIDADE FEDERAL DE TOCANTINS

A Social Network Built from Digital Documents from the UFT Website

Gentil Barbosa (1), David Nadler Prata (2), Rogério Nogueira de Sousa (3), Rafael Murta (4), Elencarlos Soares Silva (5)

(1) Universidade Federal do Tocantins, Brasil, gentil@uft.edu.br
(2) ddnprata@uft.edu.br (3) rogerio@uft.edu.br (4) rafael.mansilha@gmail.com
(5) elencarlos@uft.edu.br

Resumo

Esta pesquisa propõem a exemplificação de um mapeamento de uma rede social. Desata forma, nesta pesquisa, reproduziu-se uma rede social de entes relacionados à Universidade Federal do Tocantins (UFT) por meio do uso de 13 mil documentos digitais. Para mapear as conexões dos documentos e gerar o grafo, foi utilizado como critério que caso o nome de duas pessoas estejam em um mesmo documento eles possuem uma conexão, para cada arquivo diferente que houver conexão, esta relação é fortalecida. O grafo apresentou 114.405 vértices, e 21.081.984 arestas. Dos dez nós de maior centralidade, os entes encontrados ocuparam, ou ainda ocupam as seguintes funções: reitor, diretor de campus, pró-reitor e vice-reitor. Dessa forma, considera-se a reprodução de uma rede complexa de relacionamentos, com a utilização de documentos digitais como uma alternativa viável para o mapeamento de interações sociais de núcleos que provavelmente não estão mapeados pelas redes sociais online convencionais. Assim, o grafo criado por esse modelo de rede social, construída a partir de documentos digitais de texto, apresenta uma alternativa para mapear relações entre entes, podendo ter diversas finalidades.

Palavras-chave: Redes Sociais; Redes Complexas; Analise de Documentos; Grafo.

Abstract

This research proposes the exemplification of a mapping of a social network. Thus, in this research, a social network of entities related to the Federal University of Tocantins (UFT) was reproduced using 13 thousand digital documents. To map the connections of the documents and generate the graph, it was used as a criterion that if the name of two persons are in the same document they have a connection, for each different file that has a connection, this relationship is strengthened. The graph had 114,405 vertices, and 21,081,984

edges. Of the ten most central nodes, the entities found occupied, or still occupy, the following functions: dean, campus director, pro-rector and vice-rector. In this way, the reproduction of a complex network of relationships is considered, with the use of digital documents as a viable alternative for the mapping of social interactions of nuclei that are probably not mapped by conventional online social networks. Thus, the graph created by this social network model, built from digital text documents, presents an alternative to map relationships between entities, which can have different purposes.

Keywords: Social Networks; Complex Networks; Document Analysis; Graph.

1 Introdução

A Universidade Federal do Tocantins (UFT) é composta por: servidores técnico-administrativos, servidores docentes que juntos totalizam 2178 indivíduos ativos Presidência da República (2021), e discentes, que são aproximadamente 15 mil com vínculo ativo só em graduação. Esses indivíduos são dos mais diversos grupos econômicos, sociais e étnicos. A universidade denomina tais grupos como comunidade da UFT.

Essa comunidade, com um ambiente de entes diversificados, é uma rede de relacionamentos que vão além da relação entre aluno e professor, uma vez que, conforme o artigo 207 da Constituição Federal BRASIL (1988), as Universidades Federais devem obedecer ao princípio da indissociabilidade entre ensino, pesquisa e extensão. Considerando que estes relacionamentos são construídos no interesse acadêmico, ou profissional, a universidade não possui um mapeamento deles.

Durante a passagem de indivíduos pela UFT, é comum que ocorra o registro de seus nomes em documentos, desta forma, esta pesquisa tem por objetivo realizar o mapeamento dos referidos relacionamentos da instituição com base em documentos digitais de texto (txt, pdf, doc, docx, odt, etc.), armazenados pela instituição e quais as finalidades em que este modelo de construção de redes implica.

1.1 Redes de Relacionamento

Para Marteleto (2018), as redes de relacionamentos, ou Redes Sociais, dentro do campo das Ciências Sociais, não se limitam a uma área de estudo em específico, pois podem ser encontradas entre áreas como Antropologia, Sociologia, Economia, áreas da Tecnologia, entre outras. Entretanto, o autor define o conceito de redes sociais como uma forma de compreensão da

sociedade apoiada nos vínculos relacionais entre indivíduos, que podem levar ao reforço de capacidades de atuação, compartilhamento de conhecimento, captação de recursos e mobilização.

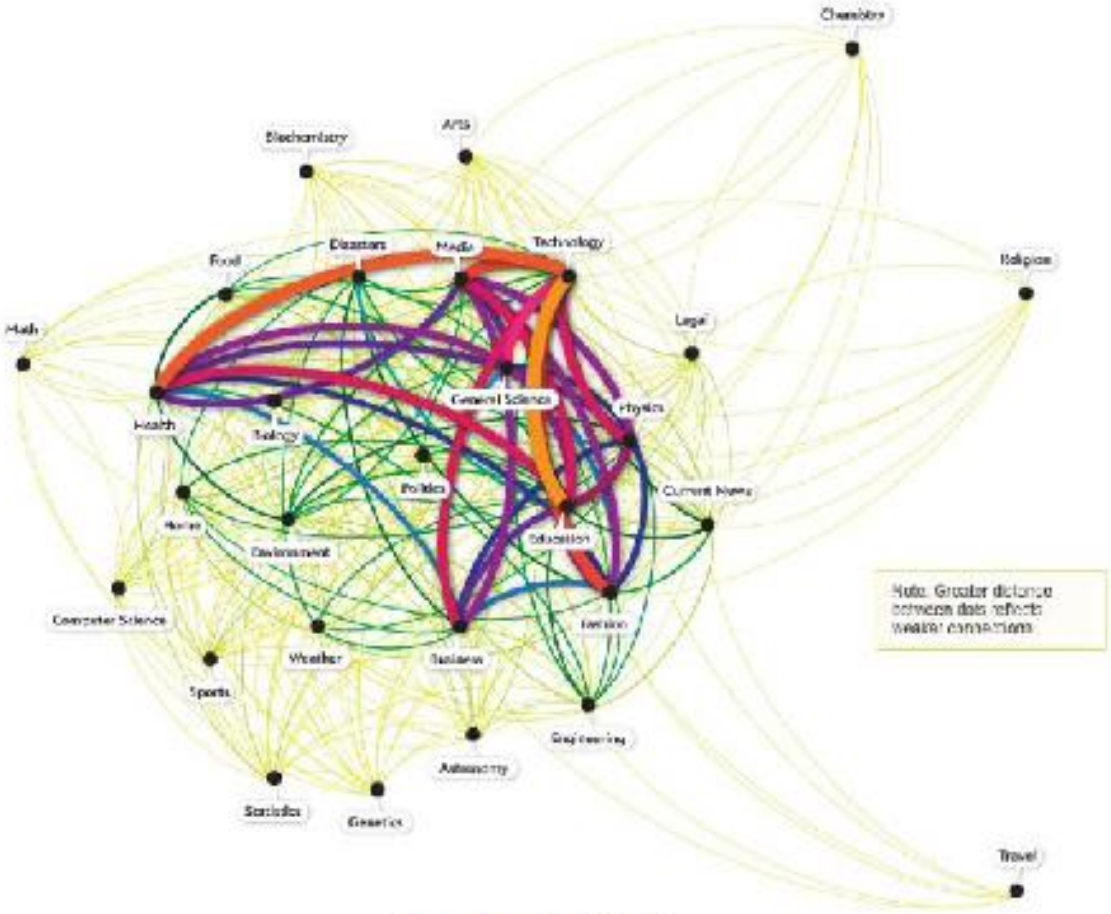
Recuero (2011) define que dois fatores são necessários para uma rede social: o primeiro fator são os atores, que constituem os nós da rede, e é composto por pessoas, instituições ou grupos; já o segundo fator é formado pelas conexões dos atores, podendo ser interações ou laços sociais, considerando assim as redes sociais como uma representação para analisar padrões de conexão de grupos sociais com base nas conexões entre os atores. Em seu estudo sobre as redes de relacionamentos na rede mundial de computadores, Recuero (2011) considera que os indivíduos e suas interações podem ser representados, de forma metafórica, matematicamente, por grafo, a representação de uma rede formada por nós (indivíduos) e arestas (suas relações).

1.2 Redes Complexas e Redes de Relacionamento

Em seu trabalho publicado na revista Scientific American, Fischetti (2011) apresentou um infográfico interativo, ilustrado na figura 1, onde um cientista da BitLy, site de encurtamento de endereço *URL*, examinou 600 páginas e rastreou 6.000 páginas que as pessoas visitaram após acessarem uma dessas 600 páginas.

4

Figura 1 – Infográfico de Tráfego na Rede de Amantes da Ciência



Fonte: Fischetti (2011)

Esse experimento construiu uma rede complexa associando os conteúdos das páginas da internet acessadas pelas pessoas. Como resultado, observou-se que havia algumas associações inesperadas, como, por exemplo, pessoas que estão interessadas em física também estão interessadas em ciência da computação, porém estas mesmas pessoas também estão muito interessadas em moda.

Inúmeros problemas do mundo real podem ser reproduzidos através de redes complexas, nas quais se conectam os nós. Existem diversos tipos de redes, como redes de cadeia de DNA, redes elétricas, redes aéreas, redes sociais (*online*), entre outras. No campo dos estudos das redes sociais, medidas de influência social vêm sendo elaboradas Newman (2005), Borgatti & Everett, A graph-theoretic perspective on centrality (2006), Borgatti, Identifying sets of key players in a

social network (2006), Agneessens, Borgatti, & Everett (2017). Em grafos como modelos para redes sociais, nós importantes são considerados mais centrais na rede Wasserman, Faust, & others (1994).

Uma rede complexa pode ser denotada matematicamente por meio de grafos. O grafo (G) é formado por dois conjuntos, sendo um de vértices, $V(G)$, que representam objetos, e outro de arestas, $E(G)$, que correspondem à relação entre os vértices Coppin (2017). As arestas podem ser identificadas por uma tupla (i,j). Em teoria dos grafos, V define a ordem do grafo ao mesmo tempo em que o número de arestas E define o seu tamanho. Diversas propriedades e características podem ser analisadas em um grafo, e as propriedades estudadas são importantes para entendimento das redes complexas.

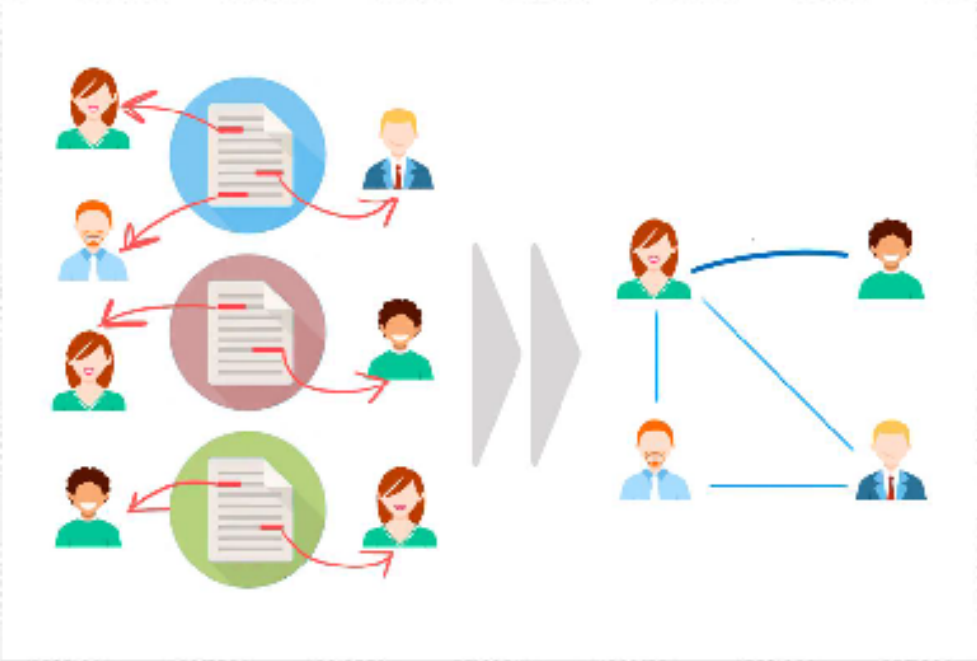
2 Metodologia

O uso do conceito de redes complexas para a construção de redes com bases de dados mostra-se uma excelente ferramenta para encontrar associações e realizar análises. Considerando que a UFT dispõe de vários tipos de repositórios para armazenamento de documentos digitais, optou-se por utilizar a base de dados de 13 mil documentos digitais, disponíveis de forma pública no portal da UFT:. www.uft.edu.br

Para a pesquisa aqui explicitada, foi considerado que, se duas ou mais pessoas estão em um mesmo documento digital de texto, então elas possuem uma relação. Esses indivíduos podem até não se conhecer de fato, mas definitivamente há uma correlação de interesse entre eles. Sendo assim, os vértices são as pessoas identificadas durante a análise dos documentos que compõem o corpus, e as arestas são formadas a partir da ocorrência de duas ou mais pessoas no mesmo documento; este processo é ilustrado na figura 2. Logo, se os nomes de duas pessoas estão no mesmo documento, elas estão diretamente conectadas.

6

Figura 2 – Processo de criação da rede complexa de relacionamentos dos entes da UFT



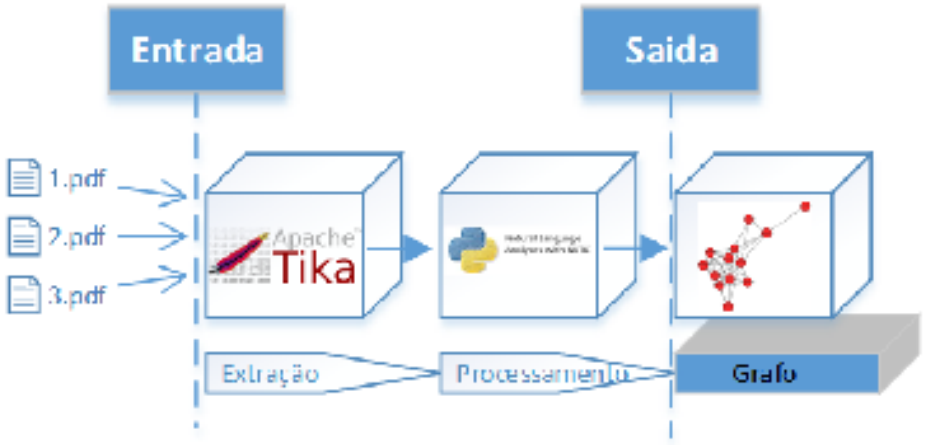
Fonte: Elaborada pelo autor

Os 13 mil documentos utilizados neste experimento foram extraídos do portal da UFT, sendo essa extração feita pela Coordenação de Desenvolvimento de Software da UFT, por meio de uma ferramenta própria de exportação da plataforma utilizada para disponibilizar o Portal. Para o experimento, foram coletados apenas documentos disponíveis de forma pública no site. Para fins de replicação deste estudo, pode-se utilizar um *web crawler*[1] para coletar os arquivos.

Neste acervo de documentos, há apenas documentos que são disponíveis para *download*, e não foram coletadas publicações ou artigos do portal. Em sua grande maioria, foram coletados documentos com as extensões doc, docx e pdf. Ao caracterizar os documentos selecionados para o experimento, os principais tipos de documentos encontrados foram normativas, resoluções, editais, manuais, instruções, relatórios, informes e panfletos. Além disso, conforme representado pela figura 3, o Apache Tika[2] foi utilizado para extrair o texto dos documentos.

7

Figura 3 – Processo de extração de nome para criação da rede complexa de relacionamentos dos entes da UFT

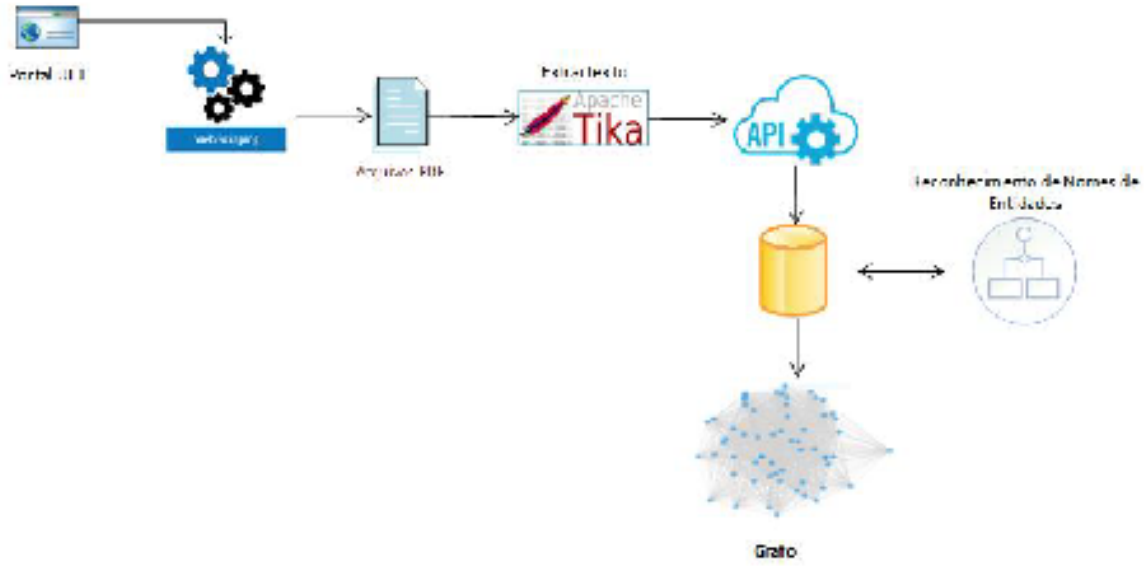


Fonte: Elaborada pelo autor

Para formação dos vértices foi utilizada a técnica de Reconhecimento de Entidades Nomeadas (REN), que se refere à tarefa de identificação e classificação de unidades de informação capaz de referenciar entidades, como nomes de pessoas, organizações, locais e datas, a partir de fontes de dados não estruturados como documentos, tarefa essa largamente utilizada no campo do Processamento de Linguagem Natural (PLN) (Nadeau & Sekine, 2007). Todo este processo é representado pela figura 4.

8

Figura 4 – Formação do Grafo



Fonte: Elaborada pelo autor

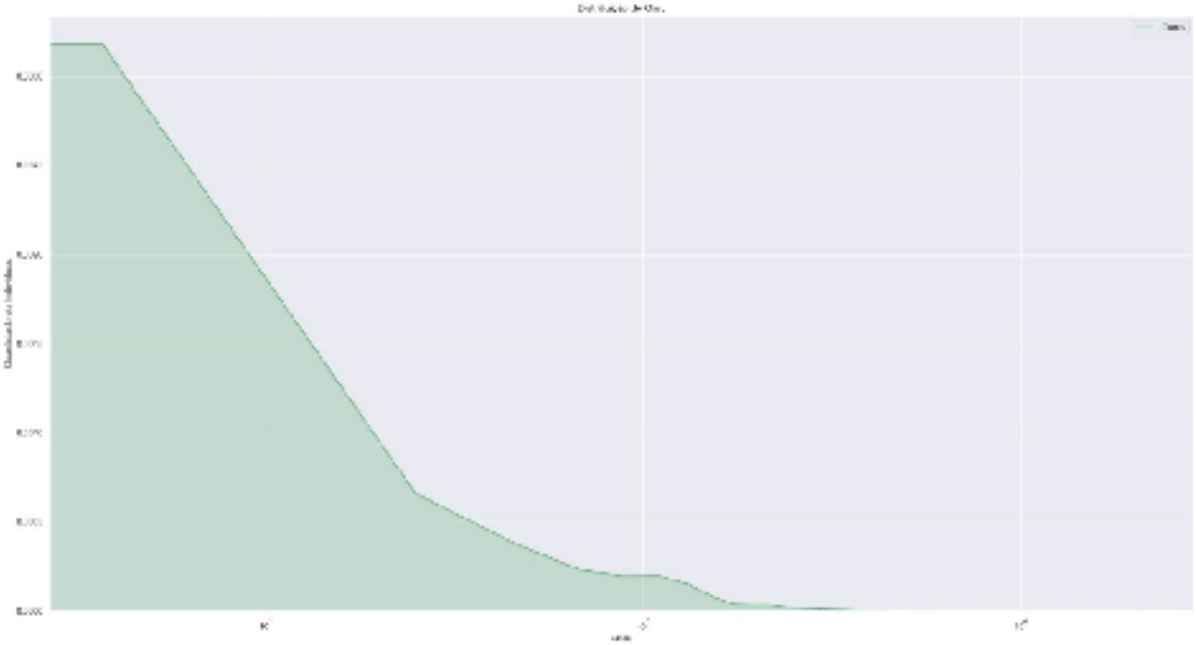
Devido à semelhança entre os documentos extraídos do portal UFT e textos jurídicos, optamos pela implementação das técnicas de REN fazendo uso do *dataset Named Entity Recognition in Brazilian Legal Text* (LeNER-BR[3]) em conjunto com o modelo de arquitetura *Long Short-Term Memory - Conditional Random Fields* (LSTM-CRF) (Araujo, et al., 2018), desenvolvido para recolhimento de entidades nomeadas contidas em textos legais em língua portuguesa. O LeNER-BR possibilita a identificação de entidades dos seguintes tipos: Pessoa, Organização, Local, Tempo, Legislação e Jurisprudência. Para este trabalho, utilizou-se apenas as palavras classificadas como Pessoa, para formação dos vértices do grafo. Cabe salientar que o LSTM-CRF (Lample, Ballesteros, Subramanian, Kawakami, & Dyer, 2016) atingiu o índice de precisão de 91.80% no que tange à classificação de entidades do tipo Pessoa ao ser treinado e testado no Corpus Paramopama (Mendonça, et al., 2015).

O peso das arestas é dado pela quantidade de documentos em que o mesmo par de indivíduos são citados, portanto, quanto mais documentos citam um par de pessoas distintas, mais conectados eles são. Dessa forma, o experimento trata da análise de uma rede complexa do tipo ponderada.

3 Resultados e Discussões

A distribuição de graus é relevante para a caracterização de uma rede. Nesse sentido, podemos observar no gráfico de distribuição de grau plotado com escala logarítmica, representado na Figura 5, a presença um pequeno grupo de indivíduos que possuem uma grande quantidade de conexão; em contrapartida, um grande grupo de indivíduos apresenta uma pequena quantidade de conexões. Os vértices do pequeno grupo com alta conectividade são denominados *hubs*[4] e ficam ao longo da cauda apresentada no gráfico da figura 5.

Figura 5 – Distribuição de Grau



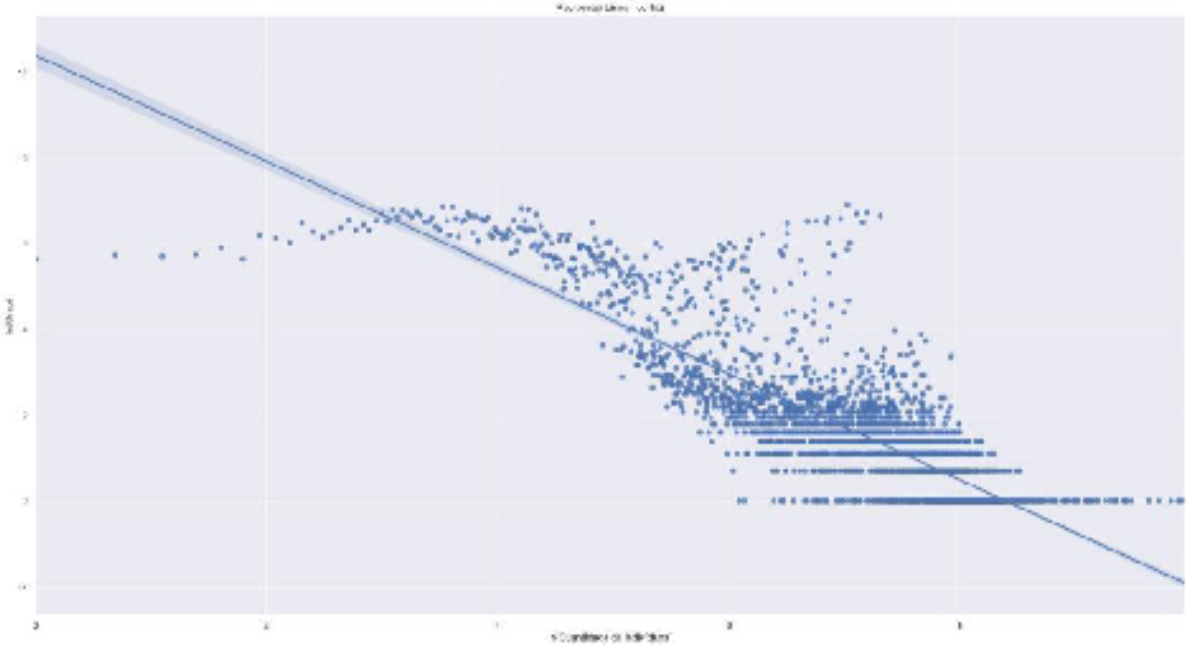
Fonte: Elaborada pelo autor

Observou-se que, ao analisar os nós altamente conectados, os 10 principais entes encontrados foram, ou são egressos e discentes que participaram de diversos editais, como processos seletivos para ingresso à universidade, auxílios estudantis, seleção de bolsas, Exame Nacional de Desempenho dos Estudantes, e outros do gênero.

Na figura 6 pode-se observar de forma mais explícita a existência dos *hubs*, e que a dispersão das frequências de grau acompanha a linha da regressão aplicada à distribuição na escala

log-log, demonstrando a relação matemática entre os escalares frequência e grau. Logo, a rede aqui estudada tende a ser uma rede de livre escala Clauset, Shalizi, & Newman (2009).

Figura 6 – Regressão linear logarítmica



Fonte: Elaborada pelo autor

O grafo gerado apresentou o número de 114.405 vértices, ou entes, com 21.081.984 arestas. O grau médio foi de 368,55, este valor alto se dá a alta conectividade dos nós do grafo. Levando em consideração que os nós são entes, em média, uma pessoa se conecta com outras 368 indivíduos, e por ser uma média alta, dado o contexto, isso implica que poucas pessoas não estão conectadas à rede.

Já a média de clusterização é de aproximadamente 0,909, isso quer dizer que, para cada nó, há aproximadamente 90% de probabilidade de um nó X conectado ao nó Y, também possuir uma aresta com outro nó Z que também está conectado ao nó Y.

A tabela 1 apresenta uma lista dos 10 nós de maior centralidade de grau dentro da rede complexa social gerada. Dos nomes apresentados, três entes ocuparam, ou ocupam, o cargo de reitores da UFT, outros três ocupam o cargo de diretor de câmpus; nove ocupam, ou ocuparam, alguma pró-reitoria e três ocuparam, ou ocupam, o cargo de vice-reitor.

11

Outra forma de aferir a importância de um vértice para uma rede complexa é medir sua centralidade autovetor Bonacich (1987), o que, assim como a centralidade de grau, leva em consideração a quantidade de nós diretamente relacionados. Essa medida também considera as ligações com nós altamente conectados, o que significa que, além do seu grau de conexão, ela calcula também a importância dos seus vizinhos.

Tabela 1- Lista dos 10 nós com maior centralidade de grau

Entidade	Cargo/Relevância	Centralidade	Grau
Indivíduo número 1	Reitor(a)	0.1811	20722
Indivíduo número 2	Pró-reitor(a)	0.1741	19923
Indivíduo número 3	Vice-Reitor	0.1712	19589
Indivíduo número 4	Reitor(a)	0.1608	18407
Indivíduo número 5	Pró-reitor(a)	0.1447	16565
Indivíduo número 6	Reitor(a)	0.1324	15154
Indivíduo número 7	Pró-reitor(a)	0.1139	13040
Indivíduo número 8	Pró-reitor(a)	0.1090	12474
Indivíduo número 9	Pró-reitor(a)	0.1077	12328
Indivíduo número 10	Pró-reitor(a)	0.1076	12312

Fonte: Dados da pesquisa

O repositório utilizado para o armazenamento dos documentos do portal, conforme servidor responsável, lotado na coordenadoria de Infraestrutura de Tecnologia da Informação da Universidade, foi implementado entre os anos de 2011 e 2012, nos últimos anos em que o Professor Alan Barbiero ocupou o cargo de reitor da UFT. Desta forma, o Grafo que resultou na rede social é um reflexo da gestão entre 2012 e 2020.

O mapeamento apresentou várias relações de professores de áreas de ensino diferente, como Ciência da Computação, Letras e Ciências Contábeis, que compunham um comitê de governança digital, mas que não usavam redes sociais, ou não possuíam estas conexões mapeadas.

Observou-se também a possibilidade de mapear grupos de trabalho da universidade de acordo com a área de atuação dos professores, e com isso detectar as possíveis conexões inexistentes, ou fracas, e definir no plano estratégico da universidade ações voltadas para melhorar a interdisciplinaridade dentro da universidade, contudo como as informações sobre as pessoas,

seus cargos na universidade, são dados que não possuem acesso público, não foi possível realizar esta análise para esta pesquisa.

Dentro das análises levantadas, mas que também estão limitadas quanto ao acesso de dados não públicos, o uso do mapeamento para encontrar pessoas com conexões fracas, ou até mesmo inexistentes, com o propósito de classificar estas ocorrências em algo que é característico das atribuições do servidor, ou se há uma questão que ocasiona esta ocorrência e assim aprimorar as relações das pessoas quem compõem a universidade.

4 Considerações Finais

Dada a relevância da indissociabilidade entre ensino, pesquisa e extensão que as universidades devem observar, conforme definido na Constituição Federal BRASIL (1988), e considerando o propósito da extensão como o desenvolvimento de ações que busquem provocar a troca de conhecimentos, podemos observar que a UFT possui relacionamentos com pessoas que vão além da sua própria comunidade, pois a universidade se relaciona com indivíduos que não estão necessariamente dentro da universidade.

Sendo assim, dentro dos 114.405 entes encontrados na rede social complexa formada por este estudo, há uma quantidade expressiva de pessoas externas à comunidade da UFT. Contudo, uma limitação desta pesquisa é que não foi possível obter acesso a todos os nomes de indivíduos desta comunidade, uma vez que estes dados não são públicos, para classificar os entes encontrados e apresentar uma análise das relações do público externo e suas relações para com a comunidade da UFT.

Tendo em vista que, se dois ou mais indivíduos estão em um mesmo documento, a rede complexa aqui construída define que há uma relação entre eles, mesmo que estas pessoas não se conheçam pessoalmente (o que evidencia uma relação de interesse mútuo entre tais indivíduos), o uso deste tipo de ferramenta pode inclusive ajudar na criação de novos relacionamentos entre indivíduos que possuem interesses semelhantes.

Assim, considerando que as redes sociais online disponíveis através da internet possuem a limitação de que as pessoas, assim como suas relações para com outras pessoas, são quem definem

estas conexões, uma vez que sua adesão é voluntária, a reprodução de uma rede complexa de relacionamentos, com a utilização de documentos digitais, mostrou-se uma alternativa viável para o mapeamento de redes sociais, permitindo o estudo de entes que, por sua vez, podem não estar mapeados nas redes sociais *online* convencionais, como, por exemplo, dois colegas de trabalho que não são amigos, e por consequência não criaram um vínculo em redes sociais online, mas possuem um convívio devido a relação profissional. Sua aplicabilidade pode ser realizada em diversos tipos de segmentos, como, por exemplo, mapear relacionamentos para descobrir associações em documentos digitais de processos penais de um determinado município.

Considera-se, assim, que é possível acompanhar a evolução de uma rede de relacionamentos, baseada em documentos, utilizando-se do período de criação dos documentos como parâmetro, mas o acesso as informações dos entes é fundamental para definir as aplicabilidade desses mapeamentos dentro do contexto social em que os arquivos que mapearam o grafo foram extraídos.

Notas

(1) Do inglês "rastreador web", um *web crawler* é um programa de computador que, de forma automatizada, acessa páginas da web a fim de coletar dados ou arquivos.

(2) O Apache Tika é uma ferramenta, escrita em Java, utilizada para extração de dados e texto de mais de mil tipos diferentes de arquivos, como os doc, docx e pdf encontrados no experimento. https://tika.apache.org

(3) Este conjunto de dados pode ser consultado pelo repositório: https://github.com/peluz/lener-br

(4) No contexto de Redes Complexas, quando nós possuem um grande número de conexões com outros nós, estes nós populares são denominados de *hubs* (Barabási & Bonabeau, 2003).

Referências

Agneessens, F., Borgatti, S. P.,and Everett, M. G. "Geodesic based centrality: Unifying the local and the global". *Social Networks*, n. 49, May 2017, pp. 12–26.

Luz de Araujo, P. H., de Campos, T. E., de Oliveira, R. R. R., Stauffer, M., Couto, S., and Bermejo, P. "LeNER-Br: a dataset for named entity recognition in Brazilian legal text". In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), Canela, RS, Brazil. Springer. 2018, pp. 313–323,

Barabási, A. L., and Bonabeau, E. "Scale-Free Networks". *Scientific American*, n. 288, 2003, pp. 60–69. doi:10.1038/scientificamerican0503-60

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. U. "Complex networks: Structure and dynamics". *Physics Reports*, n. 424, 2006, pp. 175-308. doi:https://doi.org/10.1016/j.physrep.2005.10.009

Bollobás, B., Janson, S., and Riordan, O. "Sparse random graphs with clustering". *Random Structures & Algorithms*, n. 38, 2011, pp. 269–323.

Bonacich, P. "Power and Centrality: A Family of Measures". *American Journal of Sociology*, n. 92, 1987, pp. 1170–1182. doi:10.1086/228631

Borgatti, S. P. "Identifying sets of key players in a social network". *Computational & Mathematical Organization Theory*, n. 12, 2006, pp. 21–34.

Borgatti, S. P., and Everett, M. G. "A graph-theoretic perspective on centrality". *Social networks*, n. 28, 2006, pp. 466–484.

BRASIL. Constituição da República Federativa do Brasil. *Senado Federal*, 1988.

Clauset, A., Shalizi, C. R., and Newman, M. E. "Power-law distributions in empirical data". *SIAM Review*, n. 51, 2009, pp. 661–703. doi:10.1137/070710111

Coppin, B. *Inteligência Artificial*. Rio de Janeiro: LTC, 2017.

Fischetti, M. Physics *or Fashion? What Science Lovers Link to Most: Science aficionados have odd and surprising interests*. Scientific American, 2011, https://www.scientificamerican.com/article/graphic-science-science-lovers-web-traffic/

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. "Neural Architectures for Named Entity Recognition". *arXiv:1603.01360* [cs], 2016, http://arxiv.org/abs/1603.01360

Marteleto, R. M. "Redes Sociais, Mediação e Apropriação De Informações: situando campos, objetos e conceitos na pesquisa em Ciência da Informação". *Revista Telfract*, n. 1, 2018.

Mendonça, J., Macedo, H., Bisbo, T., Santos, F., Silva, N., and Barbosa, L. "Paramopama: a Brazilian-Portuguese corpus for named entity recognition". *12th National Meeting on Artificial and Computational Intelligence* (ENIAC) 2015.

Nadeau, D., and Sekine, S. "A survey of named entity recognition and classification". *Linguisticae Investigationes*, n. 30, 2007, pp. 3–26. http://www.ingentaconnect.com/content/jbp/li/2007/00000030/00000001/art00002

Newman, M. E. "A measure of betweenness centrality based on random walks". *Social Networks*, n. 27, 2005, pp. 39-54. doi: https://doi.org/10.1016/j.socnet.2004.11.009

15

Pastor-Satorras, R., and Vespignani, A. *Evolution and structure of the Internet: A statistical physics approach*. Cambridge University Press, 2007.

Presidência da República. *Detalhamento dos Servidores Públicos por Órgão, Portal da Transparência,* 2021, https://www.portaltransparencia.gov.br/servidores/orgao?ordenarPor=orgaoSuperiorLotacaoSIAPE&direcao=asc

Recuero, R. *Redes sociais na Internet*. Porto Alegre: Sulina, 2011.

Universidade Federal do Tocantins. Resolução nº 21, de 26 de outubro de 2016. *Guia de Redação e Formatação de Comunicações Oficiais*, 2016, pp. 18.

Wasserman, S., Faust, K., & others. *Social network analysis: Methods and applications* (Vol. 8). Cambridge university press, 1994.

Watts, D. J. *Small worlds: the dynamics of networks between order and randomness*. Princeton university press, 2004.

Received: 30/09/2021 Accepted: 10/12/2022