UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS, TECNOLOGIAS E SAÚDE DO CAMPUS ARARANGUÁ
CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

Ramom Giovani Guth da Silva

CitaMetrics: Uma Ferramenta para Análise de Citações

Araranguá
2023

Ramom Giovani Guth da Silva

CitaMetrics: Uma Ferramenta para Análise de Citações

Trabalho de Conclusão de Curso do Curso de Graduação em Engenharia de Computação do Centro de Ciências, Tecnologias e Saúde do Campus Araranguá da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Engenharia de Computação.
Orientadora: Profa. Andréa Sabedra Bordin, Dra.

Araranguá
2023

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Guth da Silva, Ramom Giovani
CitaMetrics: Uma Ferramenta para Análise de Citações / Ramom Giovani Guth da Silva ; orientadora, Andréa Sabedra Bordin, 2023.
65 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Araranguá, Graduação em Engenharia de Computação, Araranguá, 2023.

Inclui referências.

1. Engenharia de Computação. 2. Redes de citação. 3. Main Path Analysis. 4. Análises bibliométricas. I. Bordin, Andréa Sabedra. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Computação. III. Título.

Ramom Giovani Guth da Silva

CitaMetrics: Uma Ferramenta para Análise de Citações

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de "Bacharel em Engenharia de Computação" e aprovado em sua forma final pelo Curso de Graduação em Engenharia de Computação.

Araranguá, 01 de Dezembro de 2023.

Documento assinado digitalmente
Jim Lau
Data: 07/12/2023 09:44:43-0300
CPF: ***.464.702-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Jim Lau, Dr.
Coordenador do Curso

Banca Examinadora:

Documento assinado digitalmente
Andrea Sabedra Bordin
Data: 06/12/2023 20:23:26-0300
CPF: ***.193.220-**
Verifique as assinaturas em https://v.ufsc.br

Profa. Andréa Sabedra Bordin, Dra.
Orientadora

Documento assinado digitalmente
Alexandre Leopoldo Goncalves
Data: 06/12/2023 21:17:07-0300
CPF: ***.740.349-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Alexandre Leopoldo Gonçalves, Dr.
Avaliador
Universidade Federal de Santa Catarina

Documento assinado digitalmente
Luciana Bolan Frigo
Data: 06/12/2023 20:59:28-0300
CPF: ***.389.799-**
Verifique as assinaturas em https://v.ufsc.br

Profa. Luciana Bolan Frigo, Dra.
Avaliadora
Universidade Federal de Santa Catarina

Este trabalho é dedicado a minha mãe, que nunca deixou de me apoiar em momentos de dificuldade e sempre me incentivou a seguir em frente.

AGRADECIMENTOS

Esta seção não poderia começar sem eu agradecer a pessoa mais importante da minha vida, a minha mãe. Ela sempre esteve presente nos meus momentos mais difíceis desde que eu me conheço por gente. Devo tudo o que eu tenho hoje à sagacidade e a perseverança dela. Foi ela quem me incentivou a não desistir do curso quando eu cogitei seriamente largar tudo.

Agradeço também o Allowme e especialmente o meu chefe Fernando Guariento, por terem me proporcionado uma experiência incrível de estágio e também efetivação à uma pessoa que não sabia nada da área. Aprendi e ainda aprendo muitas coisas novas todos os dias.

Não posso deixar de mencionar a minha orientadora, Dra. Andréa Bordin, por ter proposto a ideia deste trabalho, com o qual aprendi muitas coisas para a minha vida profissional

Por fim, mas não por isso menos importante, os vários amigos que encontrei durante este curso. O Ricardo e o Patrick, com quem compartilhei muitas aventuras divertidas. E também toda a galera do prédio onde moro, o Luan, o Ivan, o Régis, o Jadson e a Gisele, que me receberam de braços abertos quando cheguei, e sempre foram muito tranquilos e divertidos.

"Julgue seu sucesso pelas coisas que
você teve que renunciar para conseguir."
Dalai Lama

RESUMO

Métodos de análise de citações, como o *Main Path Analysis (MPA)* são importantes pois permitem aos pesquisadores identificar publicações com grande impacto, entender o histórico de desenvolvimento de um setor específico e descobrir novas tendências de pesquisa. Existem diversas ferramentas de software que permitem realizar tais análises. A análise das ferramentas mais populares revelou que todas as ferramentas são do tipo *desktop* e poucas são *open source*, o que impede contribuições da comunidade. Também, apenas uma ferramenta possibilita coleta automática de publicações de bases de dados. Não obstante, somente uma única ferramenta possui uma implantação do *Main Path Analysis*, método que encontra a sequência das publicações citadas mais relevantes em uma rede de citação. Este trabalho apresenta uma ferramenta *web* e *open source* de análise de citações com foco no método de *Main Path Analysis*, denominada *CitaMetrics*. A ferramenta permite a coleta automática de publicações de uma base de publicações abertas através de *web scraping*, auxilia o usuário na validação das publicações redundantes, calcula o caminho principal (*main path*) das publicações em uma rede de citações e o exibe em um grafo interativo. O processo de desenvolvimento da ferramenta seguiu as etapas de levantamento e especificação de requisitos, projeto, desenvolvimento e testes. A análise de um caso de uso real da ferramenta demonstrou que a ferramenta auxilia o usuário nas diversas etapas de uma análise de rede de citação identificando a sequência das principais publicações por meio do *MPA*.

Palavras-chave: Análises bibliométricas. Redes de citação. Main Path Analysis.

ABSTRACT

Citation analysis methods such as Main Path Analysis (MPA) are important because they allow researchers to identify publications with a high impact, understand the development history of a specific sector and discover new research trends. There are a number of software tools that allow such analyses to be carried out. The analysis of the most popular tools revealed that all are desktop applications, and few are open source, which prevents contributions from the community. Also, only one tool enables the automatic collection of publications from databases. However, only one tool has an implementation of Main Path Analysis, a method that finds the sequence of the most relevant cited publications in a citation network. This paper presents a citation analysis tool focused on the Main Path Analysis method, called *CitaMetrics*. The tool enables the automatic collection of publications from a database of open publications through web scraping, assists the user in validating redundant publications, calculates the main path of publications in a citation network and displays it in an interactive graph. The tool development process followed the stages of requirements gathering and specification, design, development and testing. The analysis of a real use case of the tool showed that the tool helps the user in the various stages of a citation network analysis by identifying the sequence of the main publications using the MPA.

Keywords: Bibliometric analysis. Citation networks. Main Path Analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A bibliometria é uma área da Ciência da Informação que analisa a produção científica e consiste na busca e aquisição de dados, através de técnicas estatísticas e matemáticas que descrevem e compilam aspectos da literatura e outros meios de comunicação (ARAÚJO, 2006). Dessa maneira, a bibliometria tem várias motivações: descobrir tendências emergentes, padrões de colaboração e constituintes da pesquisa, além de explorar a estrutura intelectual de um determinado domínio (DONTHU *et al.*, 2021).

Neste contexto, uma das áreas de estudo da bibliometria são as citações. A citação é uma parte formal do processo de construção e comunicação do conhecimento científico, uma vez que requer que os pesquisadores referenciem as publicações cujos conceitos e métodos inspiraram ou foram usados no desenvolvimento do seu próprio artigo (GRÁCIO, Maria Claudia Cabrini, 2020). O estudo das citações é importante por vários motivos, tais como: entender o conhecimento histórico do campo de pesquisa, auxiliar na fundamentação teórica das publicações e também criar redes de citação/conhecimento, com as quais é possível identificar novas tendências de pesquisa (GARFIELD; MERTON, 1979).

Dentre os vários métodos disponíveis para realizar análises de citações, existem: o acoplamento bibliográfico (KESSLER, 1963); a análise de co-citação (SMALL, 1973); e por fim, a Análise do Caminho Crítico (do inglês *Main Path Analysis - MPA)* (HUMMON; DEREIAN, 1989). O acoplamento bibliográfico ocorre quando dois artigos distintos referenciam uma mesma publicação em comum. A análise de co-citação é baseada no princípio de dois documentos diferentes serem citados juntos na mesma publicação. O *Main Path Analysis (MPA)* baseia-se em uma rede (grafo) de citação, na qual encontra uma sequência de publicações denotada como caminho principal. Uma rede de citação é construída a medida que os autores citam outros autores, ou seja, tais redes são construídas pela comunidade científica. Também por isso tais redes não são estáticas, mudando com a passagem do tempo a medida que crescem e novas publicações são incluídas (PRICE, 1965).

O MPA parte do pressuposto de que através de uma citação o conhecimento flui ou é transferido, no sentido citado-citante. A ideia do método é encontrar, entre todos os trabalhos que compõe a rede de citação, aqueles com maior impacto. Segundo Liu e Lu (2012), o método simplifica redes de citação complexas para apenas algumas poucas publicações, o que facilita substancialmente a análise da rede. Também explicita uma sequência com os principais desenvolvimentos de um campo científico ou tecnológico, além de destacar os principais trabalhos do domínio científico da rede (LIU; LU; HO, 2019).

A sequência de publicações ou caminho principal encontrado pelo MPA permite traçar a evolução tecnológica de uma área por meio de uma rede de citações. De acordo com Chen (2006), as reconstruções históricas dos desenvolvimentos científicos e tecnológicos são ferramentas importantes para traçar planos futuros de P&D, capturar oportunidades significativas, prever as tendências da evolução tecnológica futura, etc. Por exemplo, Yu

e Sheng (2020) analisam a trajetória de desenvolvimento da tecnologia de *blockchain*, enquanto Zhang, Ma e Liu (2020) propõem um panorama com a evolução das tecnologias de desenvolvimento sustentável utilizando o método. Por fim, Liu, Lu, Lu *et al.* (2013) elabora uma revisão de literatura, explicitando os principais trabalhos e desenvolvimentos no campo de análise por envoltória de dados, do inglês (*Data Envelopment Analysis - DEA*).

Existem várias ferramentas que se propõem a fazer análises bibliométricas que incluem redes de citação, tais como: Bibliometrix (ARIA; CUCCURULLO, 2017), VOSviewer (VAN ECK; WALTMAN, 2010), Pajek (BATAGELJ; MRVAR, 1998), Ucinet (BORGATTI; EVERETT; FREEMAN, 2002), CiteSpace (CHEN, 2006) e BibExcel (PERSSON; DANELL; SCHNEIDER, 2009).

No entanto, a pesquisa e análise das principais ferramentas destinadas a análise de citações encontrou algumas lagunas. Atualmente todas as ferramentas são aplicações *desktop*, exigindo que sejam previamente instaladas no computador dos usuários, diminuindo a flexibilidade de uso. Também, apesar da maioria das ferramentas serem gratuitas, somente a ferramenta Bibliometrix é de código aberto, o que impede contribuições da comunidade científica.

Das ferramentas citadas, apenas uma delas, o Bibliometrix, possui a funcionalidade que permite a busca e a coleta automática de dados de publicações, a qual é restrita a bases de dados científicas internacionais, como a Scopus e a Web of Science. Portanto, artigos de produção científica nacional, que não costumam estar indexados nessas bases, não são passíveis de serem coletados. Não obstante, as bases de dados supracitadas e outras bases internacionais costumam ter acesso restrito. Nas outras ferramentas é necessário coletar os dados manualmente por meio de busca e exportação desses dados, para enfim inseri-los na ferramenta de análise escolhida. Esse processo aumenta o tempo da análise e o risco de erros humanos.

Outro aspecto relacionado à coleta de publicações, é que apenas as publicações extraídas dessas bases são utilizadas para a criação da rede de citação, ou seja, as referências bibliográficas (publicações citadas) utilizadas por essas publicações não são incluídas na extração, o que deixa a rede de citação resultante menos completa. Cabe também mencionar a normalização das referências, já que muitas vezes uma referência é escrita de maneiras diferentes, o que induz erros na rede de citações e deixa a análise mais confusa. Nenhuma ferramenta estudada possui funcionalidade para normalizar as referências.

Por fim, dentre as ferramentas analisadas, apenas o *Pajek* (BATAGELJ, 2003) possui a implementação do método de MPA. Entretanto, tal ferramenta é restrita a ambientes *desktop*, não é de código aberto, e é notória a dificuldade para a sua utilização. Ainda, segundo Chen *et al.* (2022), a falta de uma implementação *open source* do MPA torna mais difícil a condução de estudos e melhorias subsequentes do método.

Diante dos argumentos expostos, entende-se que existe a oportunidade para o

desenvolvimento de uma ferramenta que permita a coleta automática de referências bibliográficas de bases abertas, possua o código aberto, esteja disponível em um ambiente *web* e que ofereça um método de análise de citação mais robusto, como o MPA. A aplicação desenvolvida pode auxiliar a realização de análises bibliométricas focadas em redes de citação, as quais poderão identificar publicações pilares e de grande relevância, além de traçar o histórico do desenvolvimento da pesquisa de diversas frentes científicas.

1.1 OBJETIVO GERAL

Desenvolver uma aplicação *web open source* para analisar redes de citação bibliográficas por meio do método *Main Path Analysis*.

1.2 OBJETIVOS ESPECÍFICOS

- Apresentar uma relação das ferramentas existentes que realizam análises bibliométricas especializadas em citações;
- Desenvolver um *web scraper* para coletar publicações de uma base de dados aberta, como a biblioteca digital SBC Open Library (SOL);
- Implementar um algoritmo para detecção e eliminação de referências bibliográficas redundantes.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado nos seguintes capítulos:

- Fundamentação teórica: apresenta os métodos de analise bibliométrica baseados em citações;
- Ferramentas relacionadas: elenca e compara as funcionalidades das principais ferramentas bibliométricas atualmente disponíveis no mercado;
- Metodologia: apresenta a metodologia utilizada para o desenvolvimento deste trabalho;
- Desenvolvimento: apresenta o detalhamento do desenvolvimento da aplicação web;
- Cenário de uso: descreve um cenário de uso da aplicação desenvolvida;
- Considerações finais: descreve os desafios encontrados durante o desenvolvimento, bem como as sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

A medida que a obra de um pesquisador cresce, o conjunto de autores referenciados em suas publicações reflete sua identidade científica, em especial, a partir daqueles autores cuja reincidência ocorra com maior frequência. Ainda, embora as citações sejam afetadas pelos laços sociais entre pesquisadores (relação citante/citado), o principal motivo da citação decorre da reconhecida relevância científica dos trabalhos citados. A quantidade de citações recebidas por um autor é evidência de sua importância ao longo de sua atividade científica (GRÁCIO, Maria Claudia Cabrini, 2020).

A análise de citações permite um maior entendimento de um campo ou tópico científico, pois permite identificar as principais publicações e autores de determinado campo científico. Ademais, segundo Demetrescu *et al.* (2020), as análises de citações podem ser usadas para identificar e aprovar promoções para autores de grande relevância.

A presente seção apresenta os principais métodos de análises bibliométricas baseados em citações: acoplamento bibliográfico, análise de co-citações e *Main Path Analysis*.

2.1 ACOPLAMENTO BIBLIOGRÁFICO

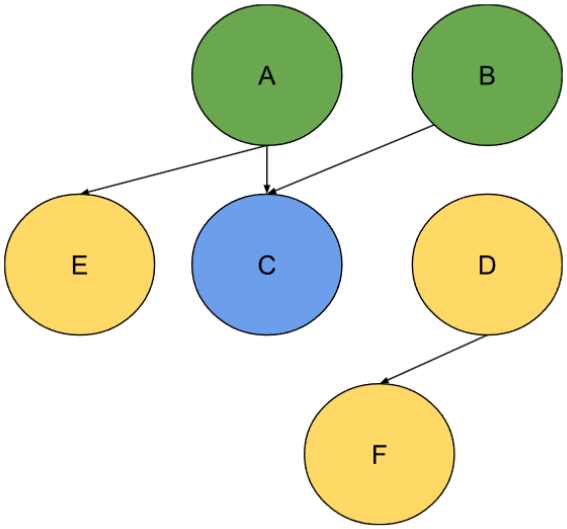
O Acoplamento Bibliográfico, também chamado de acoplamento bibliométrico, foi proposto por Kessler em 1963 e define que se há um item de referência em comum entre dois artigos, estes estão bibliograficamente acoplados, o que denota uma relação implícita entre as publicações. Quanto maior a força do acoplamento, isto é, quanto mais referências em comum dois artigos possuem, mais próximos eles estão teórico, temática e metodologicamente, portando o método indica o quanto duas publicações são similares. Tomando como base a Figura Figura 1, observa-se que A e B citam C (A,B -> C). Dessa forma, podemos inferir que as publicações A e B estão acopladas, pois ambas citam C. Os nós D e F formam um subgrafo, não conectado ao grafo principal, enquanto E não está acoplado bibliograficamente a nenhum outro nó (KESSLER, 1963).

De acordo com Carvalho (1975), tal método permite identificar o desenvolvimento de núcleos de pesquisa, assim como os artigos e pesquisadores mais importantes de um determinado campo. Ainda, segundo Jarneving (2007) o método permite particionar as publicações, criando sub-grupos, bem como permite estimar a expectativa de vida de uma publicação.

Existe uma variação, proposta por Zhao e Strotmann (2008), denominada de Acoplamento Bibliográfico de Autores (ABA), no qual dois pesquisadores são acoplados de acordo com o número de referências bibliográficas em comum que estes pesquisadores utilizam ou também pela quantidade de autores que ambos pesquisadores citam em conjunto. Apesar disso, o método é mais utilizado para analisar documentos. Também pode ser usado para fazer relações entre jornais ou instituições.

A força do acoplamento bibliográfico é fixa e não se altera com o tempo, portanto

Figura 1 – Exemplo de acoplamento bibliográfico



Fonte: Elaborado pelo autor (2023)

pode-se dizer que é uma análise retrospectiva. Entretanto o número de documentos com o qual uma publicação é bibliograficamente acoplada aumenta com o tempo, e é por essa causa que o ABA é utilizado para classificar documentos em *clusters* (MA *et al.*, 2022).

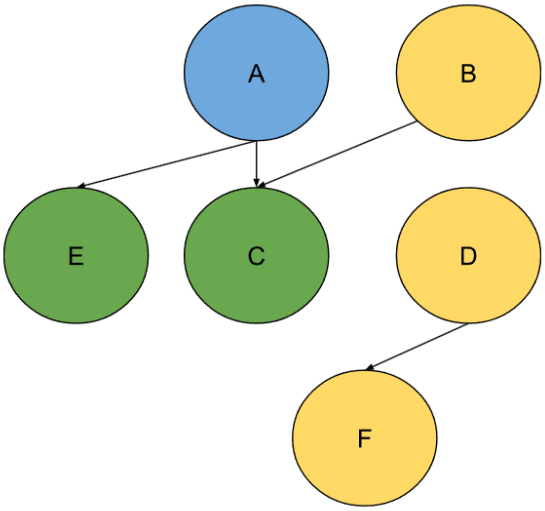
Explicitando alguns trabalhos recentes que utilizaram o método, podemos citar Lucas e Garcia-Zorita (2014), onde o conceito de Capital Social foi explorado e foram caracterizados diferentes agrupamentos de artigos de acordo com as publicações coletadas. Phan Tan (2022) buscou identificar novas frentes de pesquisa no setor de empreendedorismo social. Por fim Maseda *et al.* (2022) mapearam o envolvimento de mulheres em firmas familiares onde foram identificados diferentes núcleos temáticos, como governança e empreendedorismo.

2.2 ANÁLISE DE CO-CITAÇÃO

Proposto por Henry Small em 1973, a análise de co-citação é baseada na frequência com que dois documentos são citados juntos. Essa forma de coocorrência de dois documentos é denominada de co-citação. A co-citação identifica a similaridade de dois artigos via sua ocorrência conjunta na lista de referências de outro documento citante. Com base na Figura 2, pode-se observar que A cita C e E (A-> C,E). Dessa forma, existe uma relação de co-citação entre C e E pois ambos são citados por A. Novamente, D e F formam um subgrafo não conectado ao principal, enquanto E não possui nenhuma relação de co-citação (SMALL, 1973).

Apesar da relativa semelhança entre esses dois métodos, segundo Small (1973) e Marshakova (1981), há diferenças importantes entre os dois métodos, fornecendo padrões significativamente diferentes em relação à estrutura de um domínio científico.

Figura 2 – Exemplo de cocitação de autores



Fonte: Elaborado pelo autor (2023)

Entre as principais diferenças entre os dois primeiros métodos citados, está a interpretação dos dados, pois o acoplamento bibliográfico é considerado uma análise retrospectiva, na qual a força[1] do acoplamento é fixa. Já a análise de co-citação é prospectiva e ocorre com o passar do tempo, na medida que novos trabalhos surgem, estes citam os trabalhos antigos, portanto pode ser dito que é uma análise feita pela comunidade científica, uma vez que os próprios pesquisadores escolhem e citam os trabalhos mais relevantes (GRÁCIO, Maria Claudia Cabrini, 2020).

Uma diferença importante entre análise de co-citação e acoplamento bibliográfico é que a análise de co-citação avalia os documentos citados de uma publicação, enquanto o acoplamento bibliográfico analisa os documentos citantes de uma publicação. Entretanto, de acordo com Maria Claudia Cabrini Grácio (2020), o acoplamento bibliográfico foi pouco aplicado na análise de documentos durante sua história, onde a Análise de Co-citação se mostrou o método hegemônico. Apenas recentemente estudos que utilizam o acoplamento bibliográfico começaram a se tornar mais populares.

Para dois trabalhos serem fortemente ligados, eles devem ser citados simultaneamente por um grande número de autores. Portanto a similaridade dos trabalhos é ditada pela comunidade científica que estabelece tal conexão. Como isso é determinado pela relação dos pesquisadores, tal método evidencia como o conhecimento de determinada área científica está estruturado (MARSHAKOVA, 1981).

Por conseguinte, no momento da publicação dois artigos podem aparentar não estarem ligados, mas, segundo Maria Cláudia Cabrini Grácio (2016), com o passar do tempo essa ligação será consolidada. Como essa ligação vai ocorrendo ao longo do tempo,

[1] Entende-se "força" como a quantidade de referências em comum entre dois artigos. Quanto maior for essa quantidade, mais próximos eles estão teórica e metodologicamente.

pode-se dizer que a mesma é dinâmica e prospectiva (MA *et al.*, 2022).

Quanto mais forte for a ligação entre duas publicações, maior é a similaridade do seu conteúdo e mais próxima a relação de temas e ideias dos citados. De acordo com Hjørland (2014), a interpretação da análise de co-citação dará insumos para entender o reconhecimento e o impacto de uma publicação. Com o tempo, uma publicação pode vir a ser citada muitas vezes, tornando esta um pilar da área, logo a análise de co-citação pode identificar tais publicações.

Tal como o acoplamento bibliográfico, existe a análise de co-citação de autores (ACA), proposta por White e Griffith (1981). Neste caso, é considerado o número de artigos nos quais dois autores são citados simultaneamente. Para isso, a obra do autor citado é considerada como única e não mais a referência do artigo citado.

Alguns trabalhos que usaram a análise de co-citação são apresentados a seguir: Rafael, Herrero e Sousa (2023) analisaram o efeito placebo nas ciências sociais aplicadas, identificando as principais publicações da área; Castanha, Santos Júnior e Tolare (2023) revelaram os autores mais influentes do tema da cultura de convergência; Korte, Tiberius e Brem (2021), devido a fragmentação da pesquisa no setor de IoT (*Internet of Things*), conduziram uma pesquisa de análise de co-citação, onde encontrou e descreveu dezenove temas diferentes de pesquisa.

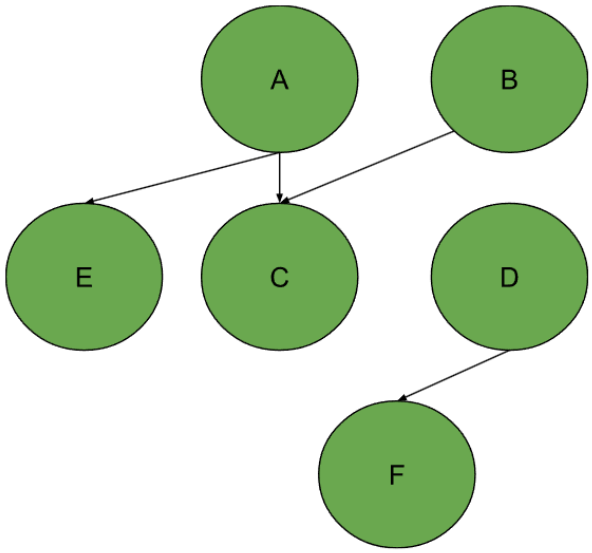
2.3 ANÁLISE DE REDE DE CITAÇÕES DIRETAS

A citação direta é a forma mais simples de se relacionar publicações e se explicita por meio da referenciação dos trabalhos, quanto um autor cita um trabalho passado no seu referencial teórico. Como pode ser observado na Figura 3, será formada uma ligação entre [A,B -> C] pois A e B citam C, outra ligação [D -> F] pois D cita F e por fim [A -> E] pois A cita E.

Em uma rede de citação, os nodos são as publicações e os links são as relações de citação para outras publicações. Em uma representação matemática, tal rede é conhecida como um grafo composto de vértices e arestas (BARABÁSI, 2013). Como existe uma direção específica, onde um trabalho pode ser citante, citado ou ambos, o grafo é direcionado. Além disso, o grafo é acíclico, pois um trabalho não pode citar a si mesmo, e salvo raras exceções, causadas por divergências nas datas de recebimento e aceite de um artigo em periódico ou jornal, só é possível citar uma publicação já existente, isto é, do passado. A essa estrutura se outorga o nome de grafo direcionado acíclico, ou do inglês, *Directed Acyclic Graph (DAG)*. A Figura 4 mostra uma rede de citação gerada pela ferramenta VOSviewer.

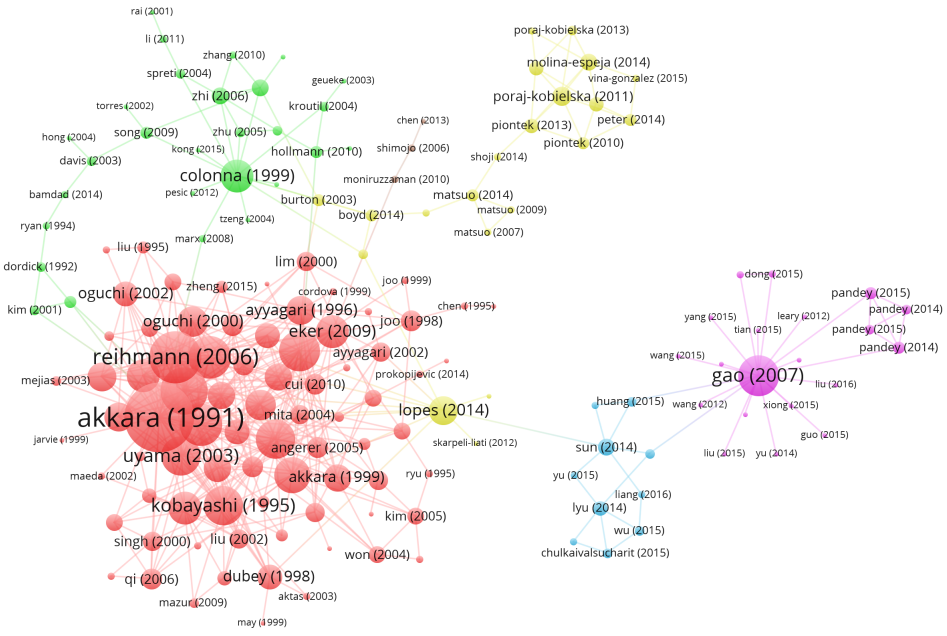
Além disso, grafos possuem algumas propriedades importantes, principalmente relacionadas ao grau, ou do inglês *degree*. O *degree* é o número total de arestas que um determinado nó possui, sendo a soma das arestas que apontam para este nó junto das arestas que partem de tal nó. O *indegree* é o número de arestas que apontam para o nó. Em

Figura 3 – Exemplo de citação direta



Fonte: Elaborado pelo autor (2023)

Figura 4 – Rede de citação



Fonte: (VAN ECK; WALTMAN, 2010)

uma rede de citação, seria o número de citações que uma determinada publicação recebe. Por fim, o *outdegree* é o número de arestas que partem do nó. Numa rede de citação, seria o número de referências bibliográficas que uma determinada publicação possui.

De acordo com Donthu *et al.* (2021), é possível analisar o impacto de uma publicação pelo número de citações que a mesma recebe. Além disso, por não depender de citações conjuntas em trabalhos distintos, essa análise é a mais rápida e adequada para analisar as

frentes de pesquisa emergentes. Entretanto, segundo Boyack e Klavans (2010), comparado com o acoplamento bibliográfico e a análise de co-citação, o método sofre com uma baixa precisão, devido ao fato que é necessário um grande volume de dados (publicações) para se identificar padrões significativos. Também, segundo Liu, Lu, Lu *et al.* (2013), redes de citações grandes podem ser complexas de serem analisadas manualmente.

O método de análise de citações diretas é utilizado nos trabalhos de Hota, Subramanian e Narayanamurthy (2020), onde são mapeados nove setores distintos de produção intelectual do setor de empreendedorismo social, identificando a perspectiva do campo, os trabalhos mais influentes bem como analisando a comunicação social dessas redes de pesquisa. Também, no trabalho de Dawson *et al.* (2014) onde é feita uma análise sobre o campo de analíticas de aprendizagem, concluindo que existe uma certa fragmentação entre as principais disciplinas do tema, bem como os trabalhos mais citados são de uma natureza mais conceitual, refletindo a necessidade dos autores de definirem o tema.

2.4 MAIN PATH ANALYSIS

Proposto por Hummon e Dereian (1989), o *Main Path Analysis (MPA)* é uma ferramenta matemática utilizada para analisar os principais caminhos do conhecimento a partir de uma rede de citações. O método proposto identifica através de diferentes métricas quais os artigos mais importantes de determinada rede de citação e o fluxo do conhecimento nesta rede, isto é, encontra a cadeia de citações mais significante e representativa da rede. Pode ser usado para mapear fluxos de conhecimento, conduzir revisões da literatura, explorar histórico de patentes, entre outros.

A partir de uma rede de citações, que pode ser considerada um grafo direcionado acíclico (do inglês DAG), o método consiste de duas etapas principais:

1. **Atribuir um peso a cada nó:** Inicialmente todos nós, e consequentemente suas arestas (citações) ou do inglês *links* da rede tem o mesmo peso de importância para o fluxo de conhecimento. O método se vale de algumas métricas para atribuir um peso a cada *link*, de forma que as citações com maior peso se tornem mais importantes na rede. A isso o autor dá o nome de *Traversal Weight* (Peso de Travessia).

2. **Encontrar o caminho principal:** Calculado o *Traversal Weight*, é necessário filtrar os nós que possuem maior relevância na rede, assim encontrando o principal fluxo do conhecimento. Essa busca pode ser feita localmente ou globalmente.

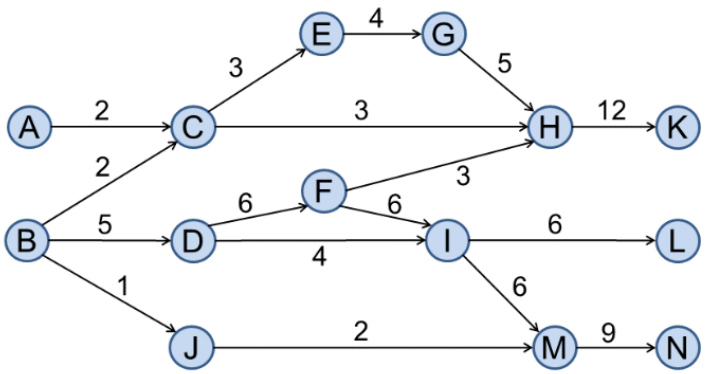
Existem vários métodos e discussões na literatura sobre qual é a melhor maneira para atribuir os pesos a cada *link*, tais como *Node Pair Projection Count* (NPPC), *Search Path Link Count* (SPLC) e por fim *Search Path Node Pair* (SPNP), todos eles propostos por Hummon e Dereian (1989). Um outro algoritmo foi proposto por Batagelj (2003), o

Search Path Count (SPC). Entretanto, segundo Liu, Lu e Ho (2019), o melhor método é o *Search Path Link Count* (SPLC), que será o método a ser explorado nesse trabalho.

Para clarificar a primeira etapa do MPA, é necessário definir alguns termos: **fontes** são nós que são citados mas não citam ninguém; **destinos** são nós que não recebem citações mas citam outros; **nós intermediários** citam e são citados; e por fim os **ancestrais** de um determinado nó são os nós que podem alcançar tal nó através da rede de citação.

Definidos os termos, a primeira etapa consiste em buscar por todos os caminhos que passam através de um *link*, incluindo os caminhos que se originam do próprio *link* e seus ancestrais, até encontrar um nó de destino. O total de caminhos encontrados é o *Search Path Link Count (SPLC)* de tal *link*. A Figura 5 representa uma rede de citações de exemplo, com os SPLC calculados.

Figura 5 – Exemplo de um grafo de Main Path Analysis usando SPLC



Fonte: Johnliu.tw - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=61755537

Tomando como exemplo o *link* G -> H, seu SPLC é 5, pois este é o número de caminhos que passam por tal *link*, que são:

1. A → C→E→G→H→K
2. B→C→E→G→H→K
3. C→E→G→H→K
4. E→G→H→K
5. G→H→K (o próprio *link*)

Outro exemplo é o *link* A -> C:

1. A→C→H→K
2. A→C→E→G→H→K

Por fim, utilizando o *link* I -> M como exemplo:

1. B→D→I→M→N
2. B→D→F→I→M→N
3. D→I→M→N
4. F→I→M→N
5. D→F→I→M→N
6. I→M→N

De acordo com Liu, Lu e Ho (2019), o SPLC é o melhor método para ser utilizado, pois é o que mais se assemelha a difusão do conhecimento no mundo real. Fazendo uma analogia de como o conhecimento é difundido, os nós iniciais criam e "repassam" o conhecimento para os nós intermediários, que por si só também criam e repassam conhecimento para os nós de destino. Alguns métodos não consideram os nós intermediários como criadores de conteúdo, o que evidentemente não se aplica num caso real. Além disso, o fato de ter um tempo de processamento razoável, em torno de N^2, onde "N"é o número de nós da rede de citação, também é um fator importante.

Por fim, na segunda parte do método, se faz necessário encontrar o caminho principal que representará o MPA. O criador do método de MPA propõe um algoritmo de busca de prioridade, o qual a partir de uma fonte escolhe localmente a próxima aresta com o maior valor de SPLC, até encontrar um nó de destino. Seu funcionamento é análogo a um algoritmo guloso e, portanto, Liu e Lu (2012) o denominam busca local. Por sua vez, Batagelj (2003) propõe o algoritmo de caminho crítico, que busca a rede inteira pelo caminho com o maior *Traversal Weight* possível, portanto Liu e Lu (2012) definem o mesmo como busca global.

Tal algoritmo é exatamente o contrário da busca pelo menor caminho. Considerando que o SPLC calculado anteriormente seja a distância entre os vértices, o algoritmo da busca global busca um caminho único na rede de citação com a maior distância possível. Dadas as propriedades da rede de citação (direcionada, cíclica), um algoritmo de ordenação topológica é capaz de encontrar o menor (ou maior) caminho no grafo (CORMEN, 2001). Por este motivo, o algoritmo de busca global foi escolhido para ser implantado nesta ferramenta.

Alguns trabalhos que utilizaram o *MPA* são apresentados a seguir: Barbieri *et al.* (2017) realizaram uma busca na base de dados da *Web of Science* sobre o tema de *Environmental Innovation* (evolução ambiental), no qual foram coletadas 2033 publicações, e posteriormente filtradas a 833 publicações, onde foram removidas aquelas que não citavam nenhuma outra publicação da base. Foi utilizado o *software Pajek* para fazer a análise. Também, como método de cálculo dos pesos do nós o SPC foi utilizado, bem como a busca global para encontrar o caminho principal, o qual possui doze publicações. Com a análise,

foram encontradas quatro sub-áreas principais. Por fim o autor apontou os principais resultados encontrados com a análise e indicou possíveis desenvolvimentos futuros.

Fu *et al.* (2019) elaboraram uma revisão sistemática do desenvolvimento da área de IoT *Internet of Things* utilizando o MPA, onde o SPC foi escolhido para fazer as examinações. Segundo o autor, devido a falta de informações na literatura para a construção das redes de citação, foi escrito um algoritmo do *Matlab* para a construção das redes, usando o número DOI como identificador de cada publicação. Os dados foram extraídos da *Web of Science*, no qual foram coletados 8913 publicações. Após criada a rede com o algoritmo do *Matlab*, a mesma foi inserida no *Pajek*. A rede resultante do MPA possui vinte artigos. Com a análise, foi possível identificar quatro diferentes frentes de pesquisa e três setores chave de aplicação do IoT, bem como a trajetória do conhecimento da área, identificando os principais países e publicações pertinentes.

Por fim, Yan, Tseng e Lu (2018) buscaram entender o motivo de veículos movidos a energias alternativas (híbridos, elétricos e hidrogênio) ocuparem uma faixa de mercado inferior a 1%. Para isso analisa o desenvolvimento da pesquisa sobre estes veículos utilizando o MPA. Foram identificados 801 publicações pertinentes na *Web of Science*. O método foi aplicado utilizando o SPC e, após construída a rede constando de 35 artigos, um algoritmo de agrupamento de artigos foi utilizado para identificar as diferentes frentes de pesquisa, onde foi obtido como resultado três frentes distintas. Concluindo, de acordo com análise realizada, o autor aponta as possíveis causas da baixa penetração destes veículos no mercado e aponta sugestões para trabalhos futuros com um número maior de publicações coletadas.

3 FERRAMENTAS RELACIONADAS

Nessa seção serão elencadas e comparadas algumas ferramentas disponíveis atualmente para a análise de citações bibliográficas. As ferramentas foram buscadas através de uma revisão da literatura em bases de dados científicas, além de uma busca geral na internet. No total foram encontradas seis (6) ferramentas relevantes utilizadas para análise de citações, a saber: BibExcel, Pajek, Ucinet, CiteSpace, VOSviewer e Bibliometrix.

No quesito de avaliação das ferramentas, alguns critérios foram considerados, tais como: as principais funcionalidades bibliométricas (quais métodos de análise), a gratuidade da mesma, o fato de ser de código aberto ou não, se é ativamente mantida, se existe uma funcionalidade de coleta de dados, quais bases de dados são suportadas e por fim o tipo de arquivo que a ferramenta requer para a importação. A seguir, a avaliação de cada ferramenta é apresentada e na Tabela 1 é exibida uma síntese comparativa das ferramentas. O símbolo + indica que existe uma versão paga com mais recursos. O símbolo * indica que é somente para visualização de análises.

1. **BibExcel**: Utilizado principalmente para fazer análises bibliométricas com dados exportados do Web of Science e Scopus, apesar de poder analisar dados de outras fontes se corretamente formatados para os padrões requeridos pela ferramenta. Foi desenvolvido por Persson, Danell e Schneider (2009). A aplicação é gratuita e está disponível para o sistema operacional Windows, entretanto não é de código aberto. Permite fazer vários tipos de análises, incluindo redes de citações e de autoria, análise de co-citação e acoplamento bibliográfico. Essa ferramenta não permite a visualização em forma de gráfico, mas pode, através de vários passos, gerar um arquivo para que uma ferramenta específica de visualização, o *Pajek* crie a visualização. A ferramenta não é mais mantida, sendo a última versão do ano de 2017. Disponível em: https://homepage.univie.ac.at/juan.gorraiz/bibexcel/

2. **Pajek**: Ferramenta *desktop* para análise e visualização genérica de vários tipos de redes grandes, de até 1 bilhão de nós, e exclusiva para o Windows. Desenvolvida por Batagelj e Mrvar (1998), escrita em Delphi e gratuita para uso não comercial. Não é de código aberto. Por ser uma ferramenta generalista, pode utilizar vários tipos de entrada diferentes, incluindo dados em formato *.txt* gerados pelo BibExcel. Portanto, pode ser utilizada para fazer análises bibliográficas incluindo redes de citações e de autoria, análise de co-citação e acoplamento bibliográfico. É a única ferramenta existente que implementa o MPA, com diferentes algoritmos para calcular os pesos das arestas bem como algoritmos para determinar o caminho crítico. Possui vários tipos de visualização, como *clusters*, grafos, vetores, partições e hierarquias, além disso permite a decomposição de redes grandes em outras menores. A ferramenta é mantida ativamente. Disponível em: http://mrvar.fdv.uni-lj.si/pajek/

3. **Ucinet:**Ferramenta específica para visualização de redes sociais desenvolvida por Borgatti, Everett e Freeman (2002). É exclusiva para o Windows e gratuita por 90 dias (*trial*), após esse período é necessário adquirir a mesma. Não é de código aberto. Muito usada para análises e visualizações de redes sociais, incluindo redes de citação. Pode receber como entrada dados gerados pelo Pajek e exibir grafo/redes de citações e de autoria, análise de cocitação e acoplamento bibliográfico e permite ver várias métricas da rede, como centralidade, densidade e proximidade. Tais dados podem ser carregados diretamente, ou por meio de um arquivo *txt* ou *excel*. É ativamente mantida. Disponível em: https://sites.google.com/site/ucinetsoftware/home

4. **CiteSpace**: Aplicação Java para qualquer computador que tenha Java instalado, desenvolvida por Chen (2006), ativamente mantida e gratuita, com o intuito de analisar e visualizar dados bibliométricos provindos do Web of Science ou com o mesmo padrão. Não é de código aberto e existe uma versão paga com mais recursos. Consegue realizar acoplamento bibliográfico, análise de cocitação, grafos de citações e redes de colaboração científica. Possui funções de visualização, incluindo um gráfico de geo-localização, baseado na localização dos autores. Disponível em: https://citespace.podia.com/

5. **VOSviewer:**Aplicação Java gratuita, com versão para *desktop* e outra *web*, ativamente mantida pela Universidade de *Leiden* (BOYACK; KLAVANS, 2010). Não possui código aberto. Tem inúmeras funcionalidades, incluindo acoplamento bibliográfico, análise de cocitação, grafos de citações e redes de colaboração científica. Possui ferramentas avançadas de visualização, como grafos e *heatmaps*, além de separação por *clusters*. Trata redes exportadas do Web of Science, Scopus, Lens, Dimensions e PubMed. A aplicação *Web* é utilizada somente para para compartilhar e visualizar as análises e gráficos, mas para realizar a análise em si é necessário utilizar a versão *desktop*. Disponível em: https://www.vosviewer.com/

6. **Bibliometrix:** Aplicação escrita em R, gratuita, constantemente desenvolvida e open source. Desenvolvida por Aria e Cuccurullo (2017). Funciona pela linha de comando em qualquer computador com o *R Studio* instalado. Existe a opção de executar a aplicação em um *browser* localmente, chamada *biblioshiny*. Por meio dessa interface que as análises e coletas de dados podem ser realizadas. Opera com dados exportados do Web of Science, Scopus, Cochrane, Dimensions e PubMed. É a única aplicação com uma funcionalidade de coleta automática de dados, compatível com as bases de dados citadas anteriormente. Além de calcular muitas métricas como *h-index*, também cria análises e gráficos de acoplamento bibliográfico, análise de cocitação, grafos de citações e redes de colaboração científica. Todas as análises possuem várias visualizações diferentes, como gráficos em estrela ou *heatmaps*.Disponível em: https://www.bibliometrix.org/home/

Tabela 1 – Comparativo de ferramentas bibliométricas

Nome	**Gratuito**	**Mantida**	**ACA**	**ABA**	**MPA**
BibExcel	Sim	Não	Em partes	Em partes	Não
Pajek	Sim	Sim	Não	Não	Sim
Ucinet	Trial	Sim	Sim	Sim	Não
Cite Space	Sim+	Sim	Sim	Sim	Não
VOSviewer	Sim	Sim	Sim	Sim	Não
Bibliometrix	Open Source	Sim	Sim	Sim	Não

Cabe destacar que os dados utilizados pela grande maioria ferramentas bibliométricas são oriundos de bases de dados científicas internacionais que exportam os dados em alguns formatos, como *.bib* e *.RIS*, além de *.txt* e *.CSV*. Entretanto, esses dois últimos tipos de arquivos não possuem uma estrutura padronizada, com cada base de dados adotando um padrão arbitrário. Isso torna difícil criar uma ferramenta generalista, portanto as ferramentas existentes utilizam os padrões de algumas bases mais populares já pré-definidas no programa.

De acordo com o que foi exibido na Tabela 1, foram identificadas diversas lacunas nas ferramentas atualmente disponíveis no mercado. A primeira delas é que todas as aplicações existentes são do tipo *desktop*. Entretanto, este tipo de aplicação é pouco adequada para utilização do usuário final pois é necessário ser previamente instalada, nem sempre existe suporte para múltiplas plataformas e o trabalho só pode ser realizado de um único dispositivo. Portanto foi definido que a aplicação desenvolvida seria do tipo *Web*, implicando em grandes vantagens para o usuário, como facilidade de acesso em qualquer lugar ou plataforma, além de não ser necessário instalar algum programa previamente.

Outro ponto importante elencado pela Tabela 1 decorre do fato de que, apesar da maioria das ferramentas serem gratuitas, apenas uma delas possui o código aberto, o Bibliometrix. Devido a este fato, não é possível para a comunidade acadêmica contribuir para a evolução destas ferramentas, bem como a falta de implementações *open source* dos métodos bibliométricos torna difícil evoluções e melhorias em tais métodos, como pontuado por Chen *et al.* (2022). Tendo em vista estes fatos, a aplicação a ser desenvolvida será gratuita e totalmente *open source*.

Por fim, a revisão das ferramentas bibliométricas revelou que apenas uma delas, o Pajek, implementa o *Main Path Analysis*. No entanto, o Pajek é uma aplicação exclusivamente *Desktop*, de código fechado, notoriamente difícil de usar e com uma visualização dos grafos pouco interativa. Dessa forma, foi decidido que o método de análise a ser implementado na aplicação a ser desenvolvida seria o *Main Path Analysis*.

4 METODOLOGIA DE DESENVOLVIMENTO

A primeira etapa deste trabalho consistiu na revisão bibliográfica da literatura científica, que buscou estudar os métodos de análises de citação existentes, assim como a pesquisa sobre as ferramentas de análise de citação disponíveis, que mostrou o estado da prática. Tais atividades foram importantes para nortear o desenvolvimento da ferramenta apresentada neste trabalho.

Após, o desenvolvimento da ferramenta *CitaMetrics* foi realizada por meio de etapas clássicas do ciclo de vida de desenvolvimento de um software: levantamento e análise de requisitos, projeto de software, implementação e testes, conforme é exibido na Figura Figura 6.

Figura 6 – Fluxograma da Metodologia

Fonte: Elaborado pelo autor (2023)

Na etapa de Análise e Projeto da Aplicação, primeiramente foram descobertos os principais requisitos funcionais, não-funcionais e regras de negócio da ferramenta. O levantamento de requisitos foi realizado com uma especialista no domínio de análise de redes de citação. Na sequência os requisitos foram documentados e os aspectos do projeto

de interface, banco de dados e arquitetura do sistema foram definidos. Em seguida, a aplicação foi desenvolvida de acordo com os requisitos e tecnologias definidas na etapa anterior.

Por fim, foi feita uma validação da aplicação, com o objetivo de verificar o correto funcionamento por meio de um caso de uso real. Nesta etapa foram demonstradas as principais funcionalidades da ferramenta, seguindo o fluxo que o usuário final irá realizar, incluindo a coleta de publicações e referências bibliográficas de uma conferência científica indexada na base de publicações SBC Open Library (SOL), o processo de validação das publicações similares, a visualização da rede de citação completa e a identificação das principais publicações de acordo com o MPA.

5 DESENVOLVIMENTO

Nessa seção serão apresentados os resultados das etapas de Análise e Especificação de Requisitos, Projeto da Aplicação e Desenvolvimento.

5.1 LEVANTAMENTO, ANÁLISE E ESPECIFICAÇÃO DE REQUISITOS

O primeiro passo foi fazer um levantamento dos requisitos funcionais, não-funcionais e regras de negócio que deveriam ser contemplados pela ferramenta CitaMetrics. Os requisitos foram levantados por meio de entrevistas com a especialista no domínio.

A seguir são elencados os principais Requisitos Funcionais (RF) da ferramenta:

- RF1. O usuário deve se logar na ferramenta;
- RF2. O usuário pode criar um ou mais projetos de análise de citação;
- RF3. O usuário pode abrir e deletar projetos existentes;
- RF4. O usuário pode utilizar um web scraper para coletar publicações da base de artigos científicos da SBC Open Library (SOL);
- RF5. O usuário pode carregar publicações científicas por meio de leitura de arquivos CSV, com um formato pré-definido;
- RF6. O usuário deve conseguir tratar referências bibliográficas similares;
- RF7. O usuário deve conseguir visualizar a rede de citação validada;
- RF8. O usuário deve conseguir executar o método MPA;
- RF9. O usuário deve conseguir visualizar o resultado do método MPA;
- RF10. O usuário deve conseguir interagir com as redes geradas.

As Regras de Negócios (RN) de cada RF são as seguintes:

- RN1. O usuário só pode utilizar qualquer funcionalidade após efetuar o login (Associada a todos os RFs);
- RN2. É obrigatório ter ao menos um projeto criado para utilizar as opções de coleta de publicações, validar referências e visualização de grafos (Associado aos RF2, RF3, RF4, RF5, RF6);
- RN3. É obrigatório que o usuário colete as publicações através do web scraper da ferramenta ou via inserção de arquivo CSV (Associado ao RF4 e RF5);

- RN4. O MPA só pode ser executado após o usuário validar todas as referências (Associado ao RF8);
- RN5. O usuário deve ser capaz de retomar o trabalho a qualquer momento sem perda de progresso (Associado aos RF2, RF6, RF7 e RF8);
- RN6. Deve haver um fluxograma para guiar o usuário;
- RN7. Deve haver um controle interno sobre qual etapa do processo o usuário se encontra, assim como um sistema que redireciona e alerta o usuário caso este tente pular alguma etapa do fluxo pré-definido;
- RN8. Informações complementares a rede/grafo, como os nós com os maiores *indegree* e *outdegree*, devem ser exibidas para o usuário (Associado ao RF7, RF8, RF9).

E por fim os seguintes Requisitos Não Funcionais (RNF) podem ser citados:

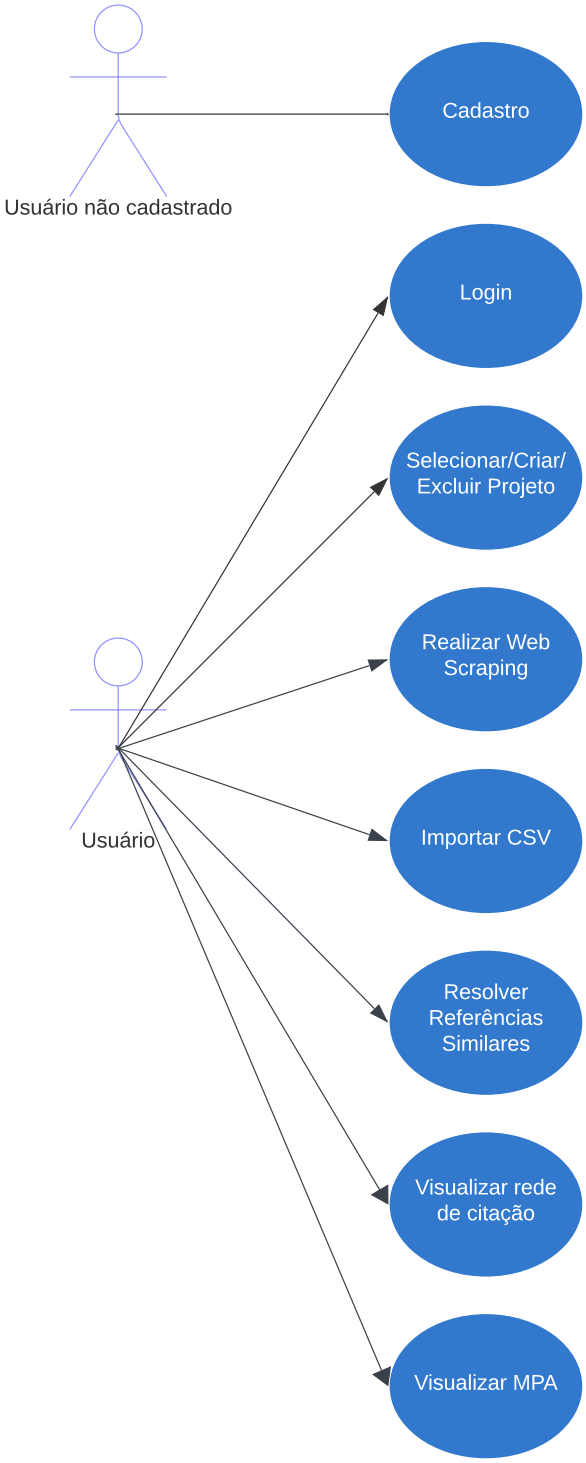
- RNF1. A aplicação inteira será desenvolvida para a web;
- RNF2. Será oferecido suporte para os browsers mais modernos;
- RNF3. Um sistema de criptografia para as senhas dos usuários;
- RNF4. A aplicação juntamente com o seu banco de dados será armazenada e distribuída em um contêiner Docker.

Na próxima sub-seção serão apresentados o diagrama de casos de uso e as descrições dos casos principais.

5.1.1 Casos de Uso

A partir dos requisitos definidos, elaborou-se o Diagrama de Casos de Uso, com o intuito de demonstrar as interações entre os usuários e as funcionalidades da ferramenta. A Figura 7 apresenta o Diagrama de Casos de Uso do sistema. O fluxo de realização dos casos de uso é explicado a seguir.

Figura 7 – Diagrama de casos de uso



Fonte: Elaborado pelo Autor (2023)

Existem dois tipos de usuário: o usuário não logado, que deve primeiramente acessar o caso Cadastrar, para usufruir de qualquer funcionalidade disponível; o usuário já logado,

que poderá interagir com todos os outros casos (funcionalidades) da aplicação, que são descritos a seguir:

- **Criar projeto**: Todo usuário precisa de ao menos um projeto para utilizar a aplicação, sobre o qual todas as outras funcionalidades irão se basear e estarão atreladas. Para este fim, existe uma página específica para criar os projetos e uma tabela para selecionar ou excluir os projetos.

- **Realizar Web Scraping**: É necessário inserir dados de referências bibliográficas de publicações científicas, ou seja, as referências dos trabalhos citados por essas publicações. Uma das maneiras de realizar esta tarefa é utilizando o *web scraper* da aplicação, desenvolvimento exclusivamente para a biblioteca digital da SBC (Open Library). O único *input* do usuário é a URL do evento/periódico a ser analisado. A aplicação irá coletar e salvar automaticamente todas as publicações encontradas na URL, bem como as respectivas referências.

- **Importar CSV**: Alternativamente, é possível inserir publicações na aplicação através do envio de um arquivo CSV com uma estrutura pré-definida. Os dados serão salvos normalmente pela aplicação.

- **Resolver referências similares**: Devido ao fato de que uma publicação pode ser citada por várias outras publicações, a sua referência bibliográfica irá aparecer diversas vezes no conjunto de referências coletadas. Além disso, essa publicação pode ser escrita de maneiras diferentes. Dessa forma, esta funcionalidade auxilia o usuário a resolver estas duplicações e inconsistências, mostrando em uma tabela todas as publicações que são similares e repetidas. O usuário então, seleciona quais são similares e escolhe uma publicação cujo formato é o mais adequado, e que representará unicamente este conjunto de publicações similares. Quando o usuário julgar que não há mais publicações similares existe um botão para encerrar o processo.

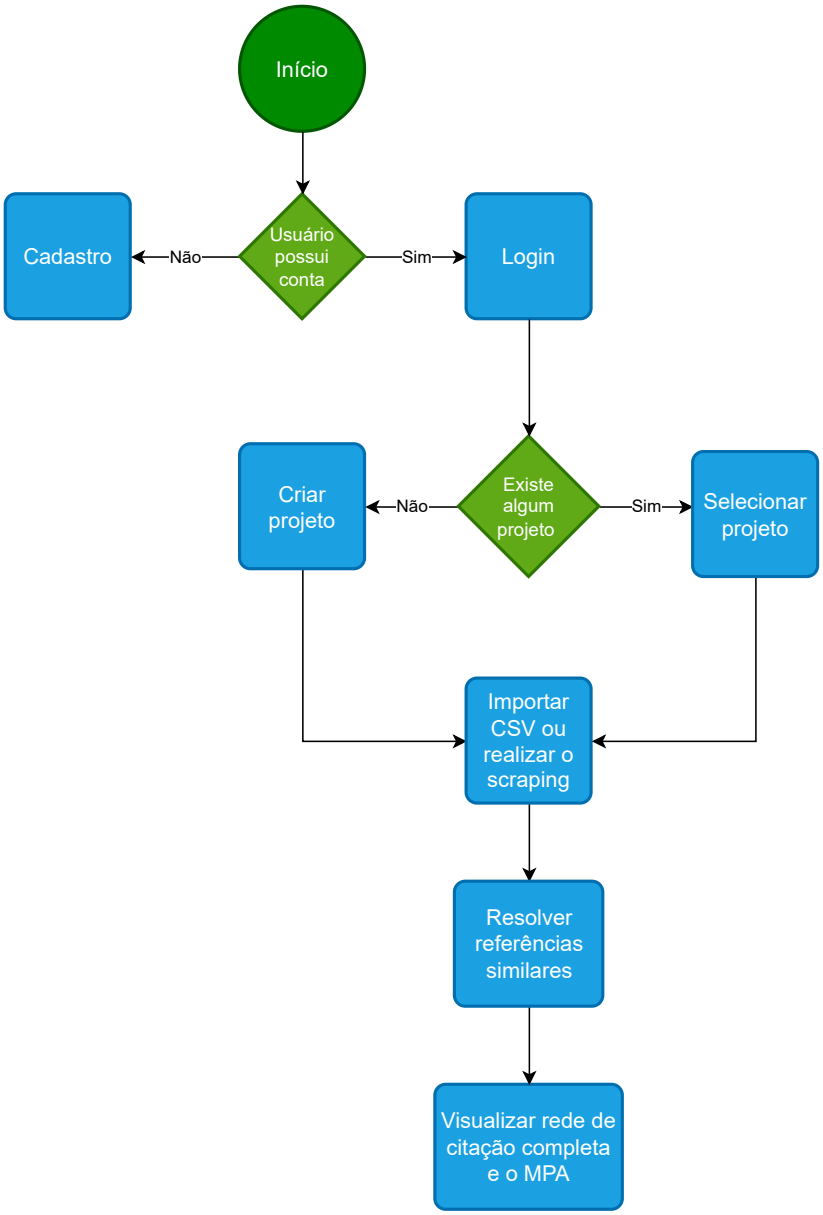
- **Visualizar rede de citação**: Também só pode ser executado após o usuário resolver todas as referências similares. A aplicação automaticamente retorna um grafo interativo com a rede de citações completa, bem como uma tabela com algumas informações úteis. Não há necessidade de interação do usuário.

- **Visualizar MPA**: Só pode ser executado após o usuário resolver todas as referências similares. Ao ser escolhida, a aplicação automaticamente calcula o MPA e exibe para o usuário o resultado em grafos interativos. Não há necessidade de outros inputs por parte do usuário.

As interações dos casos Realizar Web Scraping e Resolver Referências Similares são descritos em detalhes no Apêndice A.

Por fim, a Figura 8 mostra o fluxo de uso das funcionalidade que deve ser seguido pelos usuários.

Figura 8 – Fluxograma de uso da aplicação



Fonte: Elaborado pelo Autor (2023)

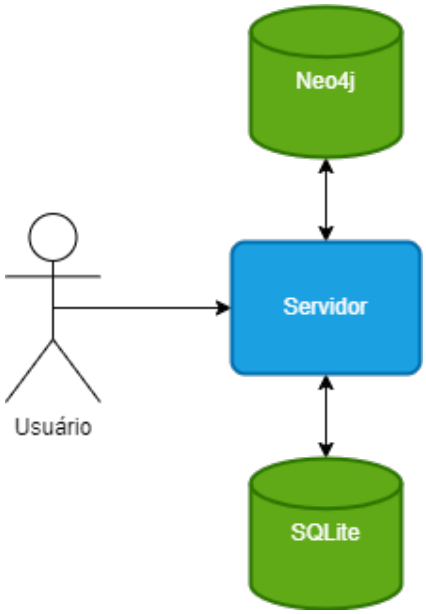
5.2 PROJETO DA APLICAÇÃO

A partir da especificação dos requisitos e casos de uso, definiu-se a arquitetura da aplicação, o projeto da interface gráfica, as ferramentas utilizadas durante o desenvolvimento, o modelo do banco de dados, o processo de resolução de referências similares e também a lógica de implementação do MPA.

A ferramenta Citametrics é uma aplicação Web tradicional, onde a arquitetura é do modelo cliente-servidor. A Figura 9 demostra uma simplificação da arquitetura da

aplicação. O propósito de cada um dos bancos de dados será explorado na sub-seção pertinente.

Figura 9 – Modelo da arquitetura cliente-servidor

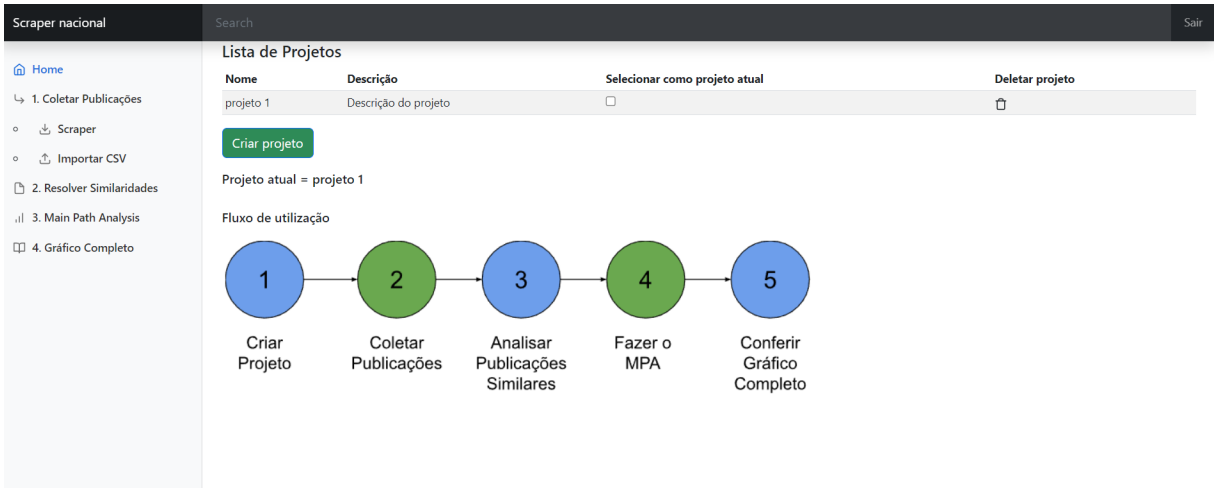


Fonte: Elaborado pelo Autor (2023)

5.2.1 Interface gráfica

No quesito de interface gráfica, esta foi projetada com um foco na simplicidade, possuindo um fluxo de uso bem definido, o que facilita a interação do usuário com a ferramenta. Para isso todos os passos são numerados e estão facilmente visíveis em um menu lateral acessível em todas as páginas. A Figura 10 mostra a página principal da aplicação, a qual possui um fluxograma para guiar o usuário, bem como a interface para criar, selecionar e deletar os projetos.

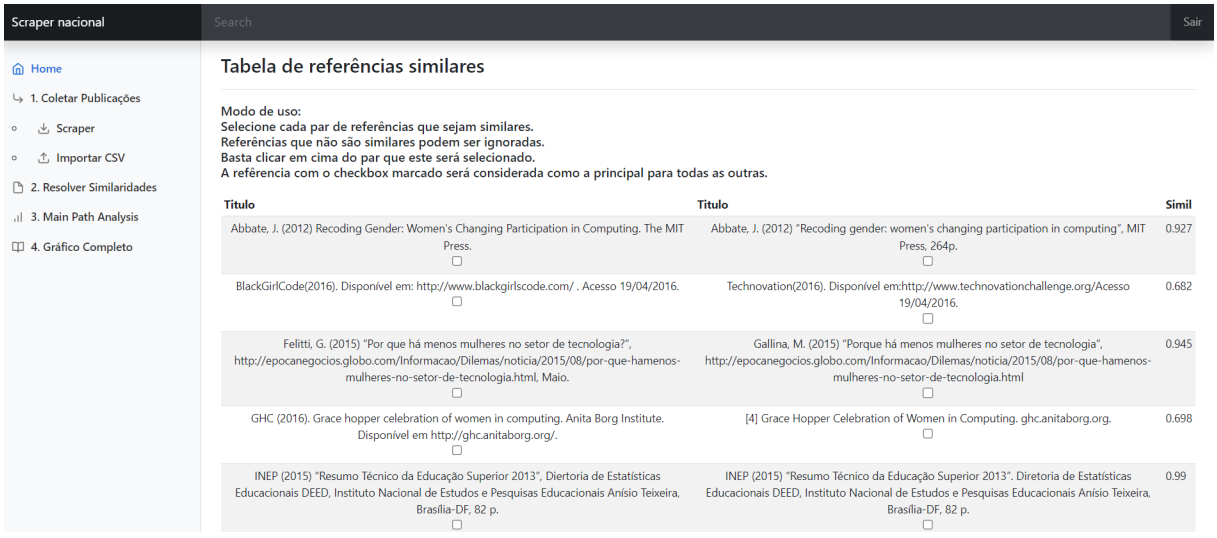
Figura 10 – Página principal da aplicação



Fonte: Elaborado pelo Autor (2023)

Na Figura 11 pode ser vista a interface para resolver o processo de similaridades. O usuário clica em todos os pares que correspondem a mesma referência/publicação, além de escolher qual será a nova publicação principal selecionando o *checkbox* desejado. Para salvar as mudanças, basta clicar no botão confirmar.

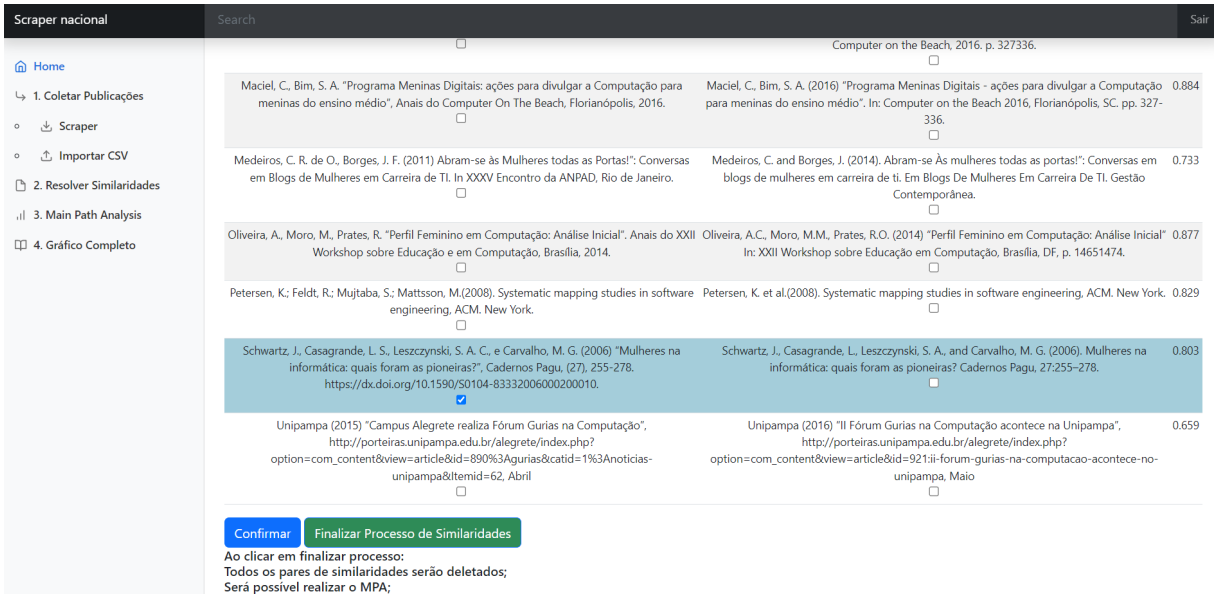
Figura 11 – Página para resolver as referências similares



Fonte: Elaborado pelo Autor (2023)

A Figura 12 mostra um exemplo da interação do usuário. O botão "Finalizar Processo de Similaridades"é utilizado para encerrar o processo, quando não restarem mais pares de publicações similares.

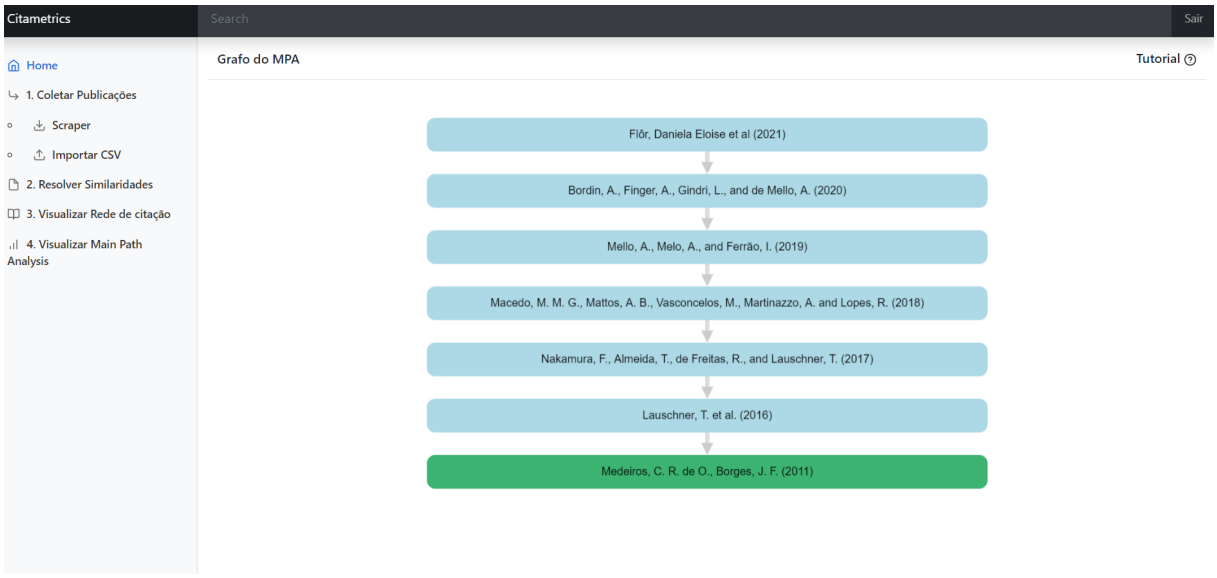
Figura 12 – Exemplo de interação



Fonte: Elaborado pelo Autor (2023)

Seguindo o fluxo, a Figura 13 mostra o grafo do caminho principal gerado pela aplicação, este processo é automático, não necessitando de nenhuma interação.

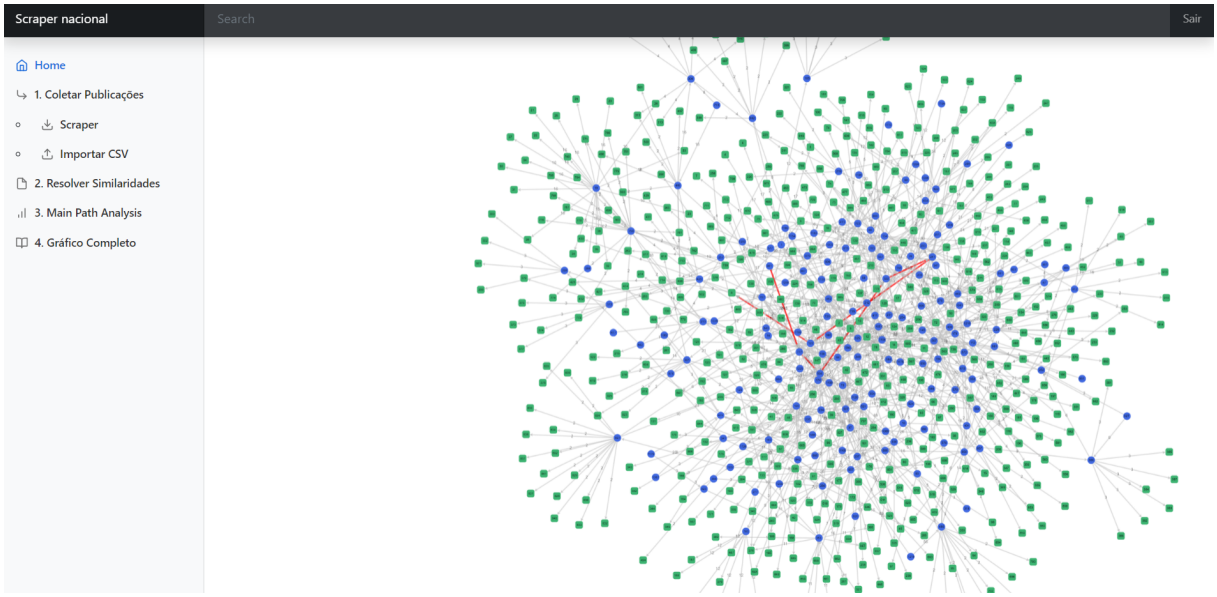
Figura 13 – Caminho principal do MPA



Fonte: Elaborado pelo Autor (2023)

Por fim, a Figura 14 mostra o grafo (rede de citação) completo com o MPA, onde as arestas vermelhas correspondem ao caminho principal. O grafo é interativo, onde o usuário pode aumentar e diminuir o zoom, mover os nós, e existe um *hover* que mostra o título da publicação ao passar por cima dos nós. O título não é exibido por padrão para uma melhor visualização do grafo.

Figura 14 – Grafo completo com o MPA



Fonte: Elaborado pelo Autor (2023)

5.2.2 Ferramentas utilizadas

Nesta seção serão especificadas quais foram as ferramentas utilizadas durante o desenvolvimento do trabalho que cumprem os requisitos especificados, e também os motivos da escolha de tal ferramenta:

1. **Python:** Esta linguagem foi escolhida devido ao fato de ser open source, possuir grande variedade de pacotes e bibliotecas existentes para a mesma, a existência de frameworks para criação de servidores *web*, uma *API* oficial para conexão com o banco de dados, grande suporte da comunidade, e por fim a sua sintaxe amigável que permite um desenvolvimento mais ágil e com um código de fácil manutenção.

2. **Neo4j:** É um banco de dados orientado a grafos muito popular, o que o torna ideal para ser utilizado neste trabalho para armazenar os trabalhos e análises, uma vez que as redes de citação criadas pelos usuários naturalmente se orientam a ter o formato de um grafo. Não obstante, é extremamente veloz e escalável, além de possuir ferramentas de visualização, muito úteis para validar a ferramenta durante o desenvolvimento.

3. **Django:** *Framework* open source para desenvolvimento de servidores *web*, sua escolha se justifica pela sua natureza flexível e escalável, que facilita o desenvolvimento de aplicações *web* robustas e de fácil manutenção. Além disso, é um framework completo, com muitas funcionalidades comuns, tais como controle de sessão de usuário e criação de *cookies*, já implementadas por padrão.

4. **Igraph:** Biblioteca open source utilizada para análise de grafos, é escrita em C e possui *ports* para Python, R e Mathematica. Implementa várias funções comuns para uso em grafos, como busca por caminhos e subgrafos. Sendo escrita originalmente em C, é muito rápida e escalável para grafos com centenas de milhares de nós. Foi utilizada para auxiliar na criação e implementação do MPA.

5. **Cytoscape:** Biblioteca open source escrita em Javascript, é utilizada especificamente para visualização de grafos interativos no *front end*, isto é, no *browser* do usuário. Dada esta especificidade, possui ferramentas únicas para grafos e, portanto, é muito adequada para este trabalho, aliada ao fato de possuir documentação extensiva.

6. **Bootstrap:** *Framework* open source para criar interfaces no *front end*, com pacotes prontos de CSS e JS, é extremamente popular e agiliza substancialmente o desenvolvimento de interfaces com o usuário, e, aliado ao fato de ser uma biblioteca open source, é muito bem documentada, o que torna a manutenção do software mais fácil.

5.2.3 Banco de dados

Na aplicação, existem dois bancos de dados, um do tipo relacional tradicional, utilizado para controle de sessão dos usuários e o outro, orientado a grafos, para todas as outras funcionalidades.

Para o controle de sessão dos usuários, tais como login e cadastro, foi utilizado o SQLite, o qual já é incluído por padrão no *framework* de servidor utilizado no desenvolvimento desta aplicação. Seu uso se justifica pela facilidade de utilização e manutenção, uma vez que não exige um servidor de banco de dados padrão pois os dados são salvos em um único arquivo. Além disso, como o volume de dados a serem salvos tende a ser pequeno e não se faz necessário o uso de consultas complexas uma vez que todas as funcionalidades do núcleo da aplicação utilizam o banco de dados não relacional, entende-se que a utilização do SQLite é adequada.

O principal banco de dados utilizado pela ferramenta é o Neo4j. Este é um banco de dados *nosql* projetado exclusivamente para o uso de grafos, portanto não existem tabelas e *joins* como em um banco relacional comum. Todos os dados salvos nele são nós, com um ou mais *labels* indicando o tipo do nó. Cada nó também possui propriedades intrínsecas, como ID ou nome, por exemplo.

Este banco foi escolhido pois as redes de citação utilizadas pela aplicação naturalmente são organizadas em grafos, logo as funcionalidades próprias deste banco tornam muito mais simples o desenvolvimento da aplicação, além disso, é extremamente performático mesmo quando bilhões de nós são armazenados.

Como não existem *joins*, para um nó se relacionar com o(s) outro(s) existem os chamados *relationships*, onde um ou mais nós se conectam aos outros. Isso nada mais é do

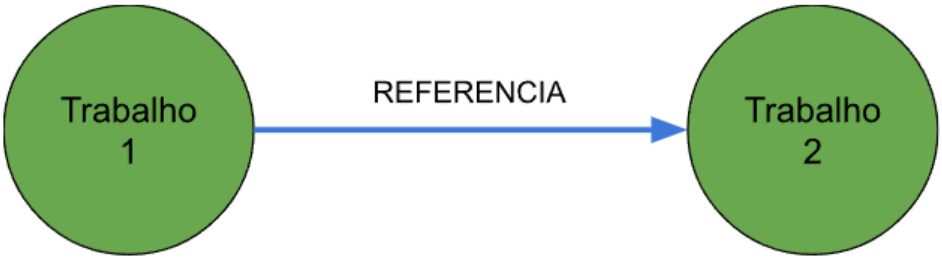
que as arestas de um grafo. Relacionamentos podem ser direcionados, ter um tipo, bem como propriedades intrínsecas, como peso ou distância.

É neste banco que ficam armazenadas as publicações coletadas de todos os usuário, bem como outras informações necessárias ao funcionamento da aplicação. Para este fim, foram utilizados quatro tipos de nó:

- **Trabalho**: Contém as publicações coletadas, bem como o tipo delas. (se cita é do tipo primário (citante) ou, se é citada por alguém, é do tipo referência);
- **User**: Contém dados do usuário, como nome e ID.
- **Project**: Contém dados do projeto do usuário. Utilizado para filtrar quais publicações serão consideradas.
- **Simil_flag**: Utilizado para controle interno sobre qual etapa o usuário se encontra.
- **URLs**: Utilizado para controle interno sobre quais URLs foram utilizadas no *web scraper*.

Tomando como exemplo os nós do tipo trabalho, uma publicação cita outra por meio do relacionamento "REFERENCIA", de acordo com a Figura 15. Também cabe mencionar que "Trabalho 1" é do tipo primário (citante) e que "Trabalho 2" é do tipo referência (citado). Neste caso, o Trabalho 1 cita o Trabalho 2, portanto este (Trabalho 2) é uma referência do Trabalho 1.

Figura 15 – Exemplo de relacionamento no Neo4j



Fonte: Elaborado pelo Autor (2023)

5.3 DESENVOLVIMENTO DA APLICAÇÃO

Nesta seção serão apresentadas as funcionalidades com desafios de desenvolvimento da aplicação CitaMetrics, entre elas o desenvolvimento do *web scraper* e importação do CSV, a resolução de referências bibliográficas similares, o do algoritmo de MPA e, por fim, o *deploy* da mesma.

5.3.1 Web Scraping e Importação de arquivos CSV

Uma das principais funcionalidades da aplicação CitaMetrics é a coleta automática de referências bibliográficas de publicações, inicialmente compatível com o site da base de dados da SBC (SOL). Para este fim, foi construído um *web scraper*, isto é, um script que coleta automaticamente as referências. Este *script* funciona através de duas etapas.

Na primeira etapa, é necessário como único *input* do usuário a URL do evento/periódico no ano/edição desejada, onde estão listadas todas as publicações pertinentes ao ano/edição do evento. A partir deste ponto, o *script*, envia uma requisição para a URL provida pelo usuário, a qual retorna o código fonte da página escolhida. Com esta resposta, o código é analisado, onde é feita uma busca por tags HTML específicas que contém as informações necessárias. A tag em questão é uma *div* com a classe *"title"*. Nesta classe existe o *link* para a publicação em si, além do título da mesma. A partir deste ponto, é criado um vetor de URLs das publicações indexadas na página.

A partir do vetor criado anteriormente, a segunda etapa consiste de um *loop* dentro desse vetor, onde para cada URL é feita uma nova requisição. Na resposta da requisição existem dois itens importantes: A lista de referências da publicação, se estas existirem, isto é, se foram incluídas no corpo da página; O outro item é um *plugin* específico para citações, no qual existe a citação completa e formatada da publicação em questão. Coletadas as informações, é feita uma limpeza nos dados para remoção de tags HTML e espaços em branco desnecessários. A partir deste ponto os dados são salvos em um banco de dados apropriado através de uma função generalista, a qual permite que trabalhos futuros expandam a aplicação implementado scripts de *web scraping* para outras bases de dados.

O *web scraping* é uma excelente técnica para coleta automática de dados quando não existe um *plugin* ou API específicas que permitam tais coletas. Entretanto, também existem desvantagens e limitações. A principal desvantagem é que um script de *web scraping* só funcionará no site para qual foi projetado, isto se não houverem inconsistências e divergências entre páginas do próprio site, logo possui alta especificidade. Não obstante, caso a estrutura do site mude, o script pode parar de funcionar, o que demandará recursos para manutenções.

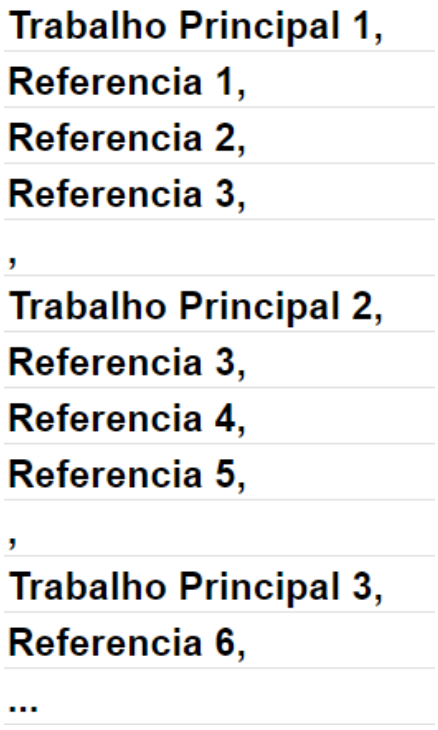
Além disso, técnicas tradicionais de *web scraping* só conseguem coletar informações que estejam disponíveis no corpo HTML da página em formato de texto simples. Caso estas informações sejam geradas dinamicamente, ou por exemplo, se o *link* da publicação retorna um documento PDF, o *web scraping* se torna substancialmente mais oneroso e difícil de implementar. Por fim, um dos desafios encontrados nesta implementação foi a indisponibilidade das referências bibliográficas das publicações, que nem sempre estavam disponíveis para coleta na biblioteca SOL. Há casos onde um ano específico de um evento possui as publicações com suas referências no corpo da página, enquanto em outros anos este corpo está vazio.

Uma outra opção disponível para a inserção de dados é através do upload de um ou

mais arquivos CSV. Isso permite uma maior flexibilidade para o usuário, pois é possível a inserção de publicações não disponíveis na base de dados da SBC (SOL). Para esta finalidade, existe uma página específica para o upload dos arquivos. Tal página também demostra para o usuário como deve ser a estrutura dos arquivos.

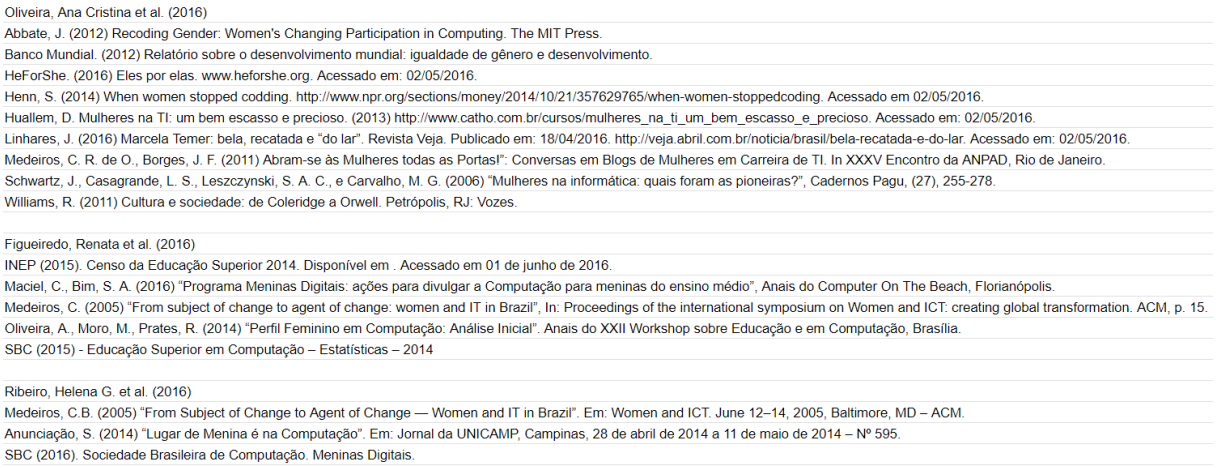
A aplicação espera como input um arquivo CSV com estrutura específica. Tal estrutura é muito simples, seguindo a ordem de Trabalho principal seguido de sua lista de referencias. Para separar um trabalho de outro, é inserido uma quebra de linha com apenas uma vírgula. A Figura 16 demostra a estrutura que o arquivo deve seguir. Caso o usuário acabe incluindo mais de uma vírgula em sequencia acidentalmente, a aplicação trata este caso, ignorando as vírgulas extras, sem exibir erros para o usuário. Já a Figura 17 demostra um arquivo populado com dados reais.

Figura 16 – Estrutura interna do arquivo CSV



Fonte: Elaborado pelo Autor (2023)

Figura 17 – Arquivo CSV com dados reais



Fonte: Elaborado pelo Autor (2023)

5.3.2 Resolução de Referências Similares

Quando a opção resolver referências similares é acessada pela primeira vez, a aplicação faz uma análise combinatória, comparando todas as publicações com todas as outras (N para N) e caso o percentual de similaridade entre um par de publicações seja maior que o valor pré-definido de 0.475 um novo relacionamento do tipo *similar_to* é criado entre esse par de publicações. Entretanto quando a comparação é feita entre duas publicações do tipo primário (citantes), a mesma é ignorada, uma vez que marcar as duas como similares significa dizer que ambas publicações distintas são iguais, o que claramente não é verdade.

Com relação ao percentual de similaridade pré-definido de 0.475, este número foi escolhido após inúmeros testes com diferentes conjuntos de dados, de maneira que o maior número possível possível de referências bibliográficas similares fossem agrupadas em pares. Este número é relativamente baixo devido ao fato de que as referências podem ser escritas de maneiras muito diferentes. Algumas citações incluem informações extras muito extensas, como data de acesso e *links* do evento/trabalho, enquanto outras não. Portanto, para tratar esse caso é necessário um percentual de similaridade mais baixo.

Feita essa comparação, os relacionamentos criados, do tipo *similar_to* serão exibidos para o usuário, que terá a decisão final em informar se esse par de publicações realmente são a mesma produção científica ou não. Caso existam vários pares que se referem a mesma publicação, o usuário pode selecionar todos. Por fim o usuário escolherá uma publicação, que será a principal e única da base, substituindo as outras escolhidas.

O algoritmo que define a similaridade entre duas publicações utiliza a *string* do nome das mesmas, e é baseado na distância de *Damerau-Levenshtein* (DAMERAU, 1964). Se o resultado for igual a 1, as duas publicações têm exatamente o mesmo nome, e se for 0 as strings são completamente distintas. Esta métrica foi utilizada pois oferece uma

performance muito satisfatória, visto que as principais variações em como uma citação é escrita se deve na maneira como é escrito o nome dos autores e dos eventos/simpósios, ou se são incluídos dados como data de acesso e onde o trabalho se encontra disponível. Por outro lado, informações como título e ano são sempre as mesmas, logo as citações possuem informações em comum, como exemplificado abaixo:

- **SILVA, JOAO (2023)** Uma publicação científica (SBC).
- **S,J(2023)** Uma publicação científica (Sociedade Brasileira de Computação).

Para melhorar a eficácia do método, quando o algoritmo faz a comparação entre duas publicações, toda a *string* da mesma é convertida em letras minúsculas, uma vez que o algoritmo é sensível a caso e também devido ao fato de que o sentido de uma frase não é alterado se a mesma se encontra em letras maiúsculas ou minusculas.

No momento que o usuário confirma que um determinado par (ou pares) de publicações são a mesma, a aplicação monta uma lista com o ID de cada uma das publicações, as quais serão excluídas da base, enquanto os relacionamentos serão alterados para apontar para a publicação que o usuário marcou como principal.

Quando este processo é encerrado pelo usuário ao interagir com o botão de finalizar o processo de similaridades, todos os relacionamentos do tipo *similar_to* são excluídos e, a partir deste ponto, a rede de citações encontra-se apta para ser utilizada no cálculo do MPA.

5.3.3 Implementação do MPA

Antes de apresentar como o MPA foi implementado, é importante mencionar algumas propriedades intrínsecas do grafo desenvolvido e armazenado no banco de dados. Conforme visto na seção passada, quando uma publicação cita outra, existe uma direção na citação (Publicação 1 cita a Publicação 2), logo isso define uma característica importante do grafo: O mesmo é direcionado.

Outra característica importante é que o grafo criado é acíclico, isto é, não possui ciclos. Novamente usando publicações como exemplo, se existisse um ciclo, seria análogo a uma publicação P ter uma referência R (P cita R), ao mesmo tempo que R cita P. Pensando em uma linha do tempo, teríamos uma publicação que cita outra do "futuro", que ainda não existe, o que é claramente impossível.

A um grafo com estas propriedades (direcionado e acíclico) é dado o nome, do inglês, de DAG *(Directed Acyclic Graph)*. É obrigatório para o funcionamento do MPA que o grafo gerado pela ferramenta seja uma DAG, caso contrário o algoritmo não irá funcionar corretamente.

Para fazer a implementação do MPA em si, o algoritmo foi dividido em duas frentes:

1. Calcular o *Traversal Weight* (pesos) de acordo com o método escolhido pelo usuário;

2. Encontrar o caminho principal;

Na primeira etapa é necessário calcular os *Traversal Weights* da rede, que de acordo com a lógica do MPA, são o número de caminhos existentes que passam através de um determinado nó. Entretanto, o que diferencia o número de caminhos possíveis na rede de citação é o método de cálculo escolhido pelo usuário.

Para uma melhor explicação do algoritmo, é importante definir previamente alguns conceitos:

- *Sources:* São todos os nós "fontes"de conhecimento, isto é, não citam ninguém, mas são citados por outras publicações (são as referências da publicação), logo possuem um *outdegree* igual a zero;
- *Sinks:* São o contrário dos nós *sources*, onde não recebem nenhuma citação, mas citam outras publicações, portanto possuem um *indegree* igual a zero.
- Intermediários: São os nós que tanto citam quanto são citados.
- Predecessores: Em um grafo direcionado, entende-se por predecessores os nós que antecedem outro. Por exemplo, em um grafo direcionado A $\rightarrow$ B $\rightarrow$ C, os nós A e B são predecessores de C.
- *Path*: Também chamado de caminho, é uma sequencia de arestas que começam e terminam em vértices pré definidos.

Pensando em uma linha do tempo, os nós *sources* são as publicações mais antigas, que "levam"o conhecimento para os nós mais atuais, os nós *sinks*.

Definidos estes conceitos preliminares, existe a definição das maneiras para calcular os pesos que cada nó possui na rede do MPA. Nesta aplicação foram implementados dois métodos de cálculo:

1. SPC: Neste método, o peso que será atribuído a uma aresta A, é o número de caminhos que passam por A, e que exclusivamente começam em nós do tipo *source* e terminam em nós do tipo *sink*.
2. SPLC: Neste método, o peso que será atribuído a uma aresta B, é o número de caminhos que passam por B, que começam em B e em seus predecessores, incluindo nós do tipo *source*, e terminam em nós do tipo *sink*.

A Tabela 2 apresenta a implementação do SPC, enquanto que a Tabela 3 representa a implementação do SPLC:

No método SPLC, para encontrar os predecessores de um vértice, pode ser feita uma busca recursiva por todos os nós, até encontrar um vértice com *indegree* igual a zero.

Tabela 2 – Implementação do SPC

```
procedure SPC(G)
    Sinks ← all nodes with outdegree = 0
    Sources ← all nodes with indegree = 0
    all_paths ← empty array
    for each source do
        for each sink do
            paths ← Find all paths between these nodes
            all_paths ← paths
        end for
    end for
    for each Edge do
        SPC_value ← find paths in all_paths containing Edge
    end for
end procedure
```

Tabela 3 – Implementação do SPLC

```
procedure SPLC(G)
    Sinks ← all nodes with outdegree = 0
    for each edge do
        all_paths ← empty list
        predecessors ← findAllPredecessors(edge)
        for each predecessor do
            for each sink do
                paths ← Find all paths between these nodes
                all_paths ← paths
            end for
        end for
        SPLC_value ← find paths in all_paths containing Edge
    end for
end procedure
```

Como visto nos algoritmos, existe uma busca de caminhos entre os vértices, e, como mencionado anteriormente, é por isso que o grafo deve ser uma DAG, uma vez que citações devem ser unilaterais (direcionadas), logo a busca por caminhos deve sempre respeitar essa direção. No quesito de respeitar a ausência de ciclos, caso ocorre-se um ciclo, o algoritmo entraria em um *loop* infinito, e consequentemente, o número de caminhos entre dois determinados nós (e o SPC ou SPLC) também seria infinito.

A segunda parte do MPA envolve encontrar o caminho principal, também chamado de caminho crítico, de acordo com Batagelj (2003). Está etapa é análoga a busca pelo menor caminho em um grafo, onde são utilizados os valores calculados na etapa anterior como a distância entre os vértices. A diferença é de que no MPA é feita a busca pelo caminho com a maior distância possível.

Novamente é observada uma vantagem do grafo gerado ser uma DAG, pois, para

encontrar o caminho com a menor - ou maior neste caso - distância é possível utilizar um algoritmo de busca em profundidade, ou do inglês DFS, *depth-first search*). Esse algoritmo encontra o caminho em tempo linear, com pior caso de O(|V| + |A|), que é a soma do número de arestas com os vértices (EVEN, 2011).

O pseudo código representado na Tabela 4 demonstra a busca em profundidade para DAGs:

Tabela 4 – Algoritmo de busca em profundidade em DAGs

```
procedure DFS(G, v)
    label v as discovered
    for each directed edge from v to w in G.adjacentEdges(v) do
        if vertex w is not labeled as discovered then
            recursively call DFS(G, w)
        end if
    end for
end procedure
```

5.3.4 Deploy da aplicação

Para disponibilizar na internet a aplicação desenvolvida neste trabalho, foi elaborado um contêiner *Docker*. Este contêiner foi criado utilizando um arquivo *Dockerfile* bem como um arquivo *docker-compose*. Esses arquivos instruem a *engine* do *Docker* de como o contêiner dever ser montado, bem como configuram a permanência dos dados. O contêiner contém uma imagem do servidor da aplicação bem como uma imagem oficial do Neo4j, o banco de dados utilizado.

Após validar o correto funcionamento da aplicação dentro da imagem *Docker* localmente, o código da mesma junto com os arquivos *Docker* necessários foi enviada ao repositório de códigos da UFSC. A partir desse momento, o SETIC criou um domínio próprio para a aplicação assim como disponibilizou um servidor próprio para hospedar o contêiner *Docker*. A aplicação *Citametrics* se encontra disponível em http://citametrics.ufsc.br

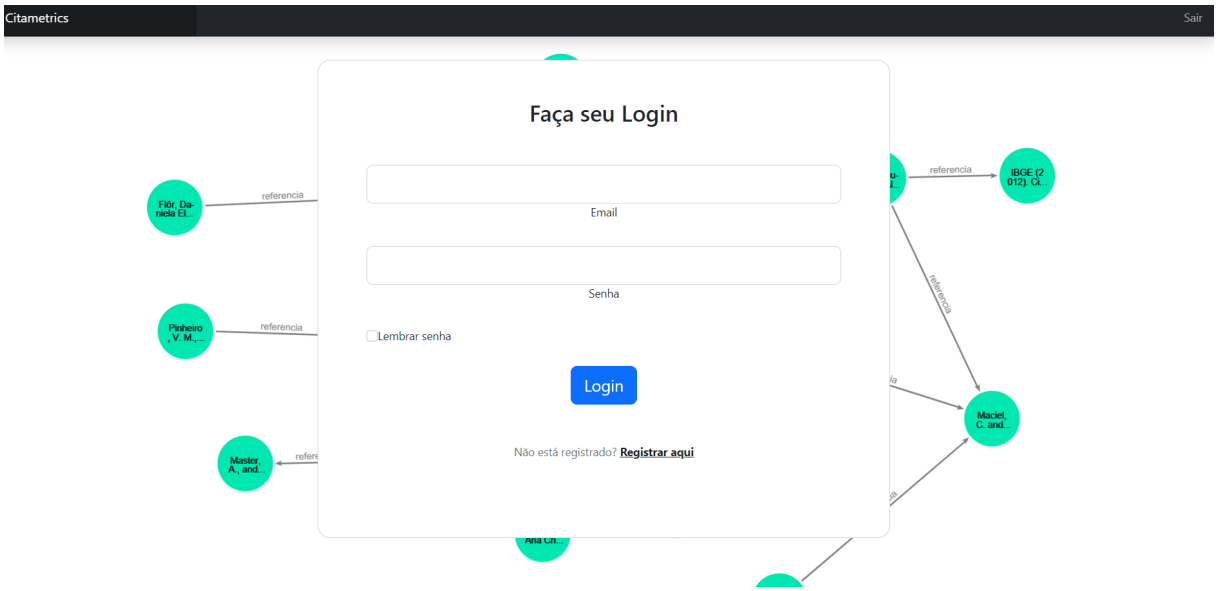
Por fim o código fonte da mesma se encontra no repositório de códigos *Gitlab* da UFSC, disponível em: https://codigos.ufsc.br/ramom.giovani.guth/mpa_analysis

6 CENÁRIO DE USO

Este capítulo visa demonstrar as funcionalidades da ferramenta de análise de citações Citametrics, por meio de um caso de uso real, desde a autenticação para uso da ferramenta, a criação de projetos, a coleta e a resolução das referências/publicações similares, a exibição da rede de citação completa e por fim a execução do MPA. Nessa avaliação, serão analisadas as publicações do Simpósio Brasileiro de Educação em Computação (EDUCOMP) considerando os anos de 2021, 2022 e 2023.

O primeiro contato do usuário com a ferramenta é através da página de *login*, exibida na Figura 18, uma vez que todas as funcionalidades são restritas a usuários logados.

Figura 18 – Página de Login



Fonte: Elaborado pelo autor (2023)

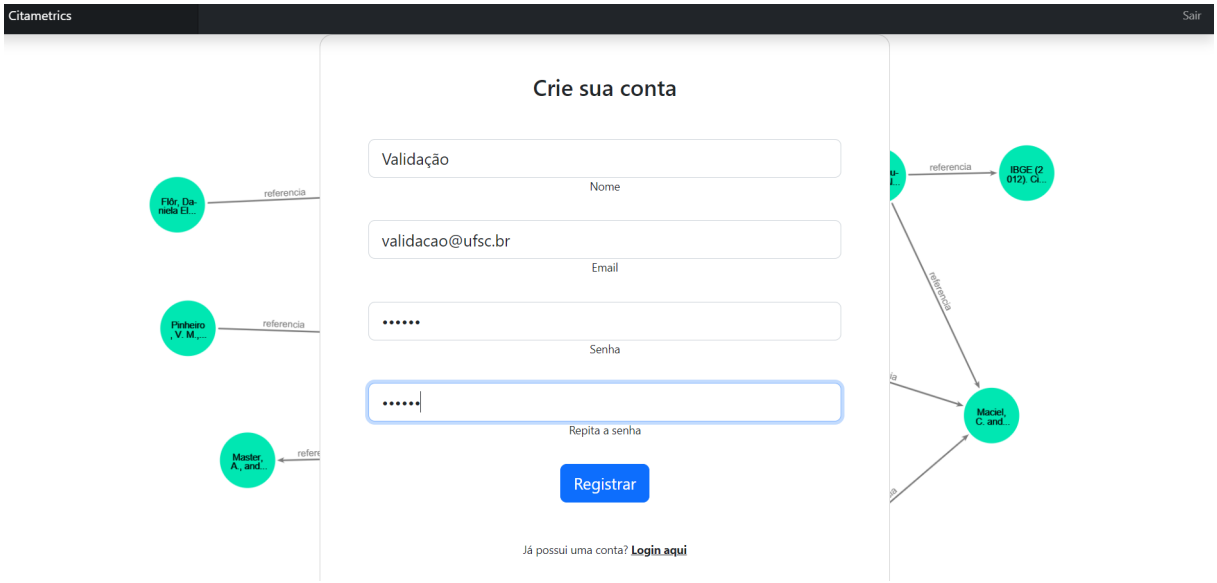
A página também contém um link que permite que novos usuários possam criar uma conta, de acordo com a Figura 19.

Efetuado o cadastro, o usuário será redirecionado para a página de gerenciamento de projetos conforme a Figura 20, onde estarão disponíveis as funções de criar, visualizar e excluir os projetos existentes. Um projeto é uma espécie de ambiente de desenvolvimento, onde o usuário pode realizar diversas análises distintas, separando as publicações e eventos por projeto.

Na página também é exibido um fluxograma para guiar os usuários, bem como o menu lateral que mostra as outras funcionalidades da ferramenta de maneira numerada para facilitar o entendimento do usuário.

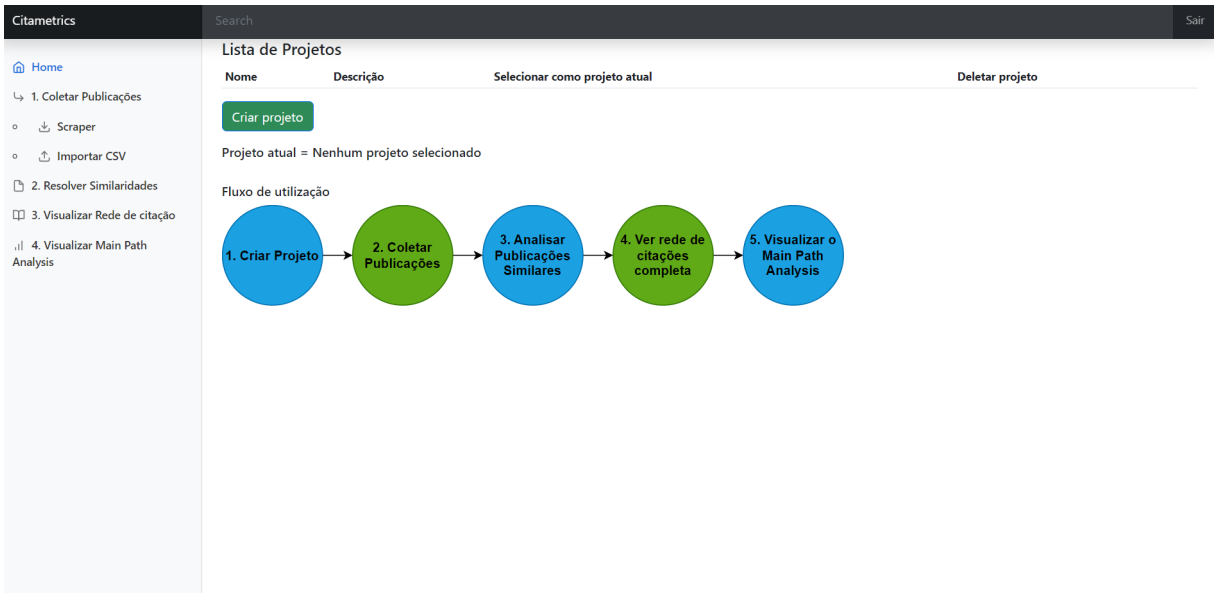
O primeiro passo do fluxo da aplicação consiste na criação de um novo projeto, como visto na Figura 21, onde deve ser inserido o nome e uma breve descrição do projeto.

Figura 19 – Cadastro



Fonte: Elaborado pelo autor (2023)

Figura 20 – Página principal

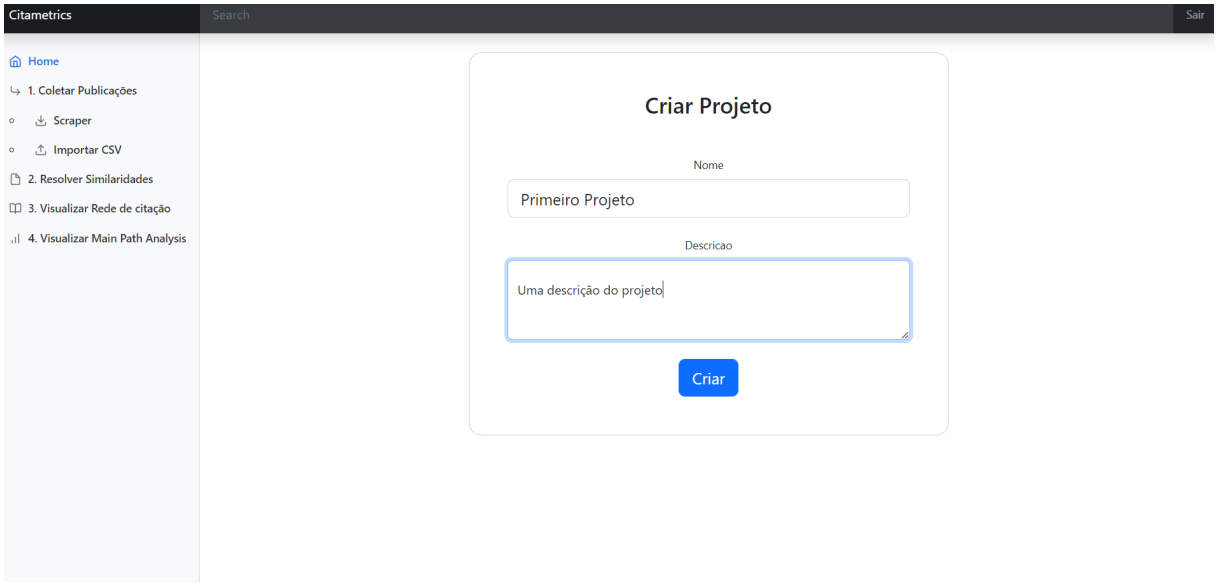


Fonte: Elaborado pelo autor (2023)

A Figura 22 mostra a página de gerenciamento de projetos com outro projeto já criado, explicitando a situação onde existem múltiplos. A tabela permite que o usuário selecione o projeto que irá utilizar, ou exclua o mesmo. Existe um *pop-up* para confirmar a exclusão do projeto, prevenindo acidentes.

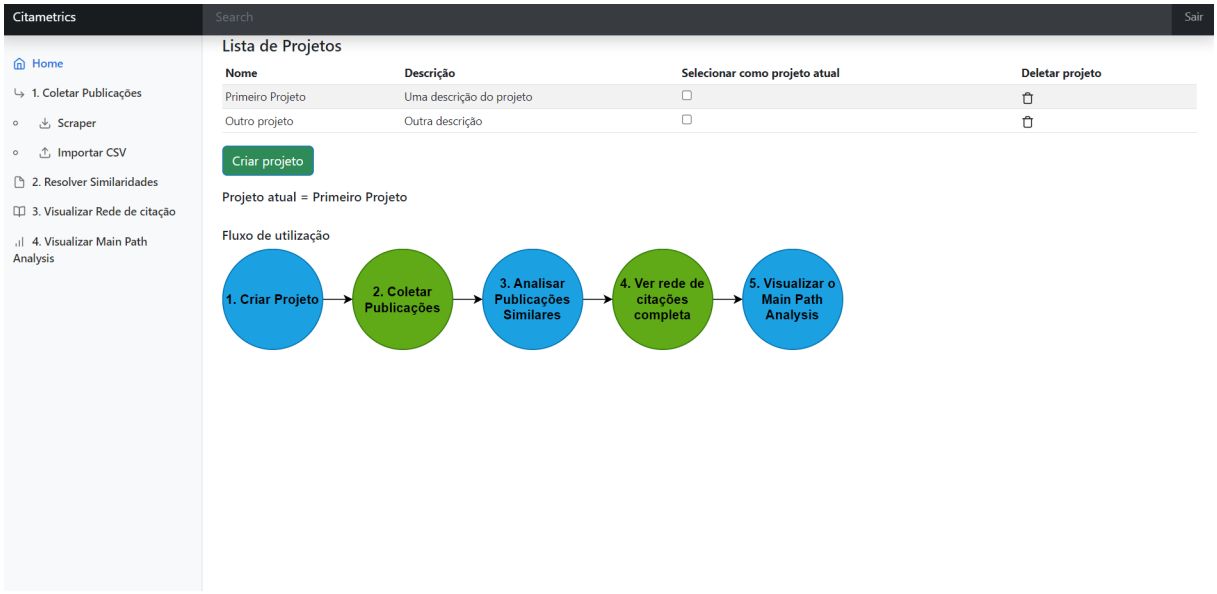
Criado o projeto, o próximo passo é a coleta ou importação das publicações. Neste caso de uso, será utilizada a coleta automática de publicações via *web scraping*, junto de

Figura 21 – Criação de um projeto



Fonte: Elaborado pelo autor (2023)

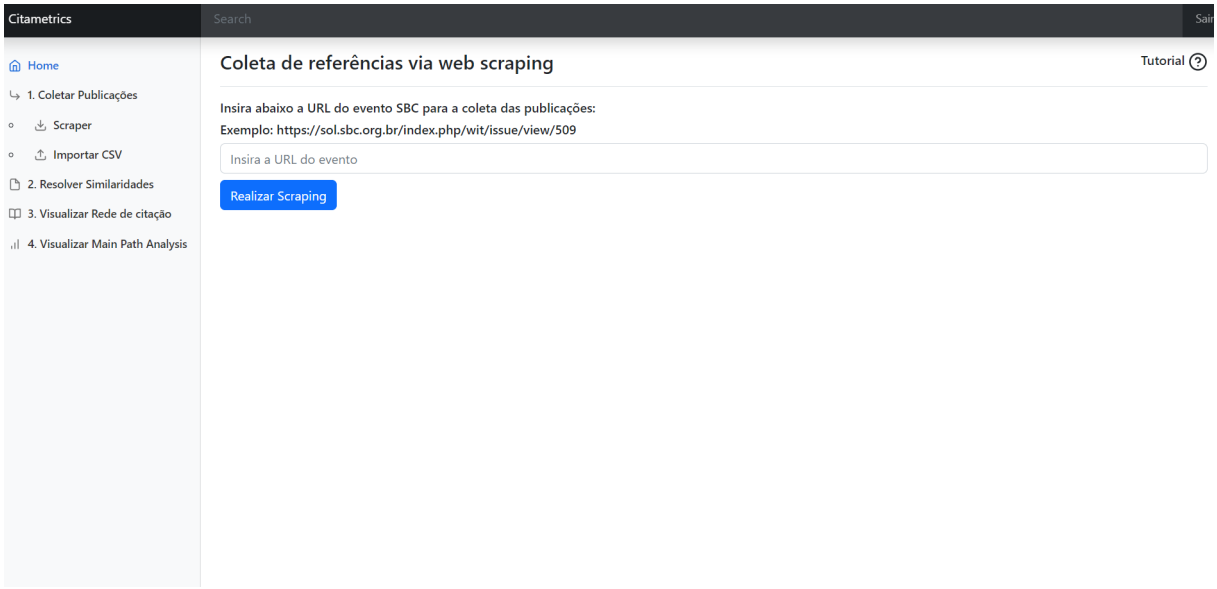
Figura 22 – Página principal com dois projetos criados



Fonte: Elaborado pelo autor (2023)

suas respectivas referências bibliográficas, conforme a Figura 23. A ferramenta, na versão atual, coleta publicações apenas da base de publicações SBC OpenLib (SOL). Desta forma, basta o usuário escolher um evento e o ano, no caso de eventos, ou o periódico e a edição para periódicos (*journals*) no qual estejam listadas todas as publicações de determinado ano ou edição. Na sequência, o usuário deve explicitamente copiar e colar a URL do evento/periódico na ferramenta CitaMetrics. A opção Tutorial, no canto superior direito, mostra algumas informações úteis.

Figura 23 – Web Scraper



Fonte: Elaborado pelo autor (2023)

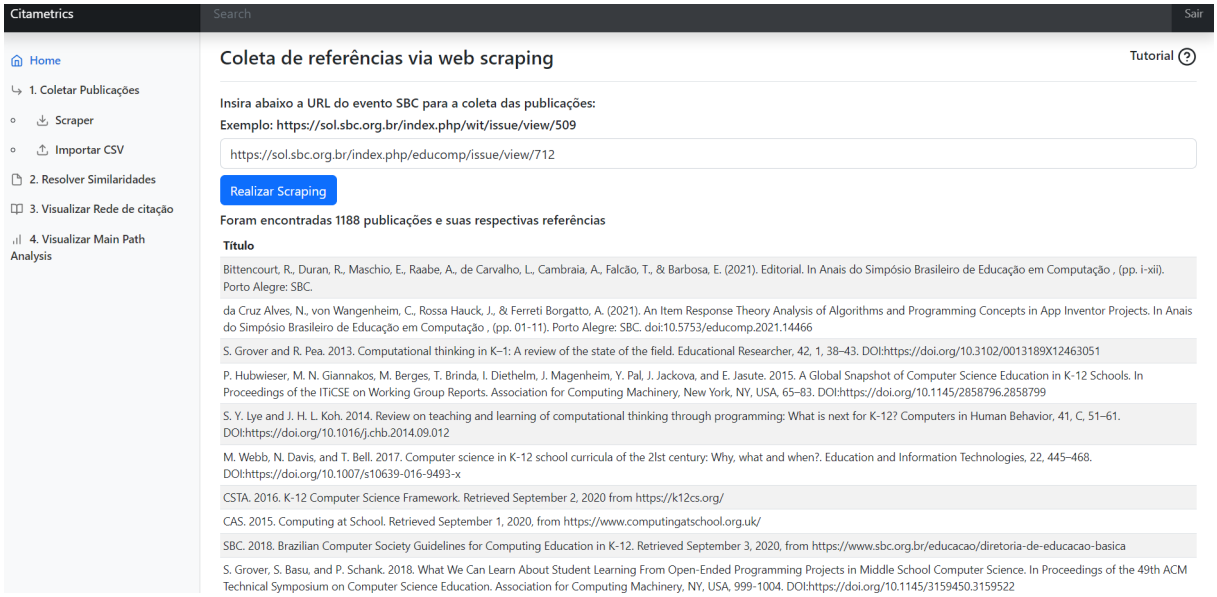
Ao indicar o *link* do evento desejado, neste caso do primeiro ano (2021) do evento EduComp, a aplicação irá coletar e salvar automaticamente as publicações e suas respectivas referências bibliográficas. Cabe ressaltar que as referências bibliográficas de cada publicação são os trabalhos citados por esta publicação. Após a coleta, as publicações e referências bibliográficas são exibidas para o usuário, conforme mostra a Figura 24, referente ao ano de 2021. Vale destacar que essa etapa pode ser repetida para todos os anos/edições da onde se deseja coletar publicações para análise, portanto, neste estudo de caso, o processo foi repetido três (3) vezes, ou seja, uma extração para cada ano. A aplicação também valida se um evento/periódico já está incluído no projeto, prevenindo duplicações indesejadas.

A Figura 25 mostra os resultados da coleta de publicações referentes ao ano de 2022, enquanto a Figura 26 corresponde aos resultados do ano de 2023.

Uma mesma publicação tende a ser citada várias vezes, entretanto, para a análise da rede de citações, é necessário que tal publicação seja representada uma única vez na base de dados. Não obstante, muitas vezes uma publicação pode ser escrita de várias maneiras diferentes. Para isso, é necessário que a ferramenta identifique as publicações similares, para que na sequência o usuário escolha, dentre um conjunto de publicações similares, qual é a publicação principal. Portanto, após a coleta das publicações das três edições/anos do evento EDUCOMP o usuário se dirige para a opção do menu Resolver Similaridades, que trata dessas questões.

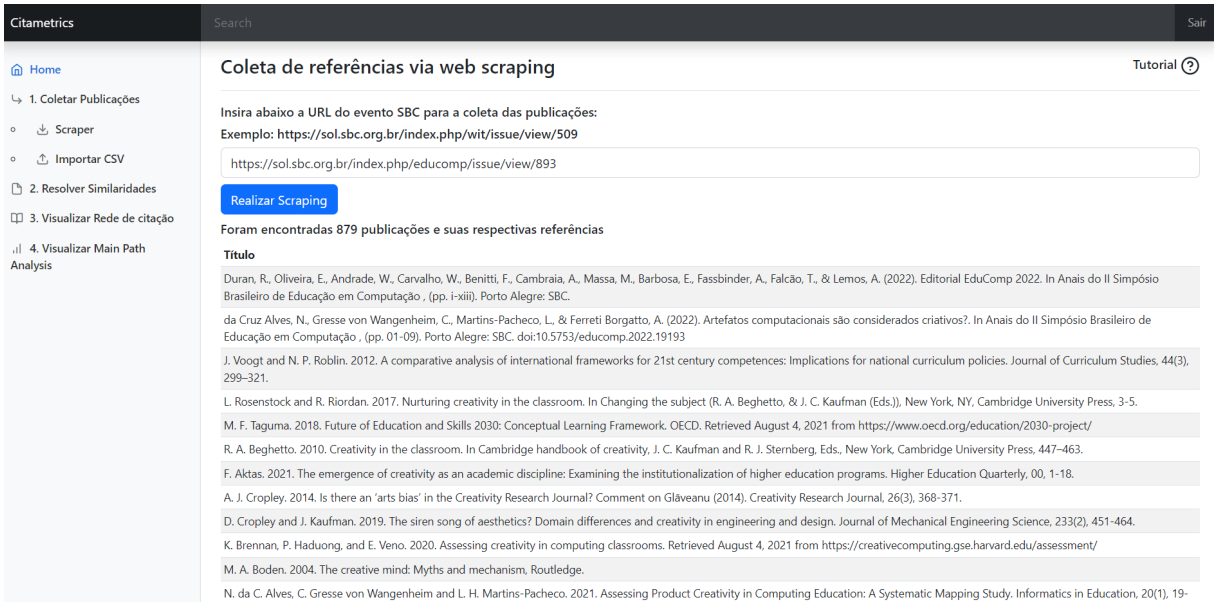
A página da Figura 27 mostra todas as publicações agrupadas par a par por similaridade. A aplicação compara as publicações e dá uma nota do quão estas são similares. Cabe ao usuário confirmar se estas publicações realmente são iguais ou não. Como pode

Figura 24 – Tabela com as publicações coletadas do ano de 2021



Fonte: Elaborado pelo autor (2023)

Figura 25 – Tabela com as publicações coletadas do ano de 2022

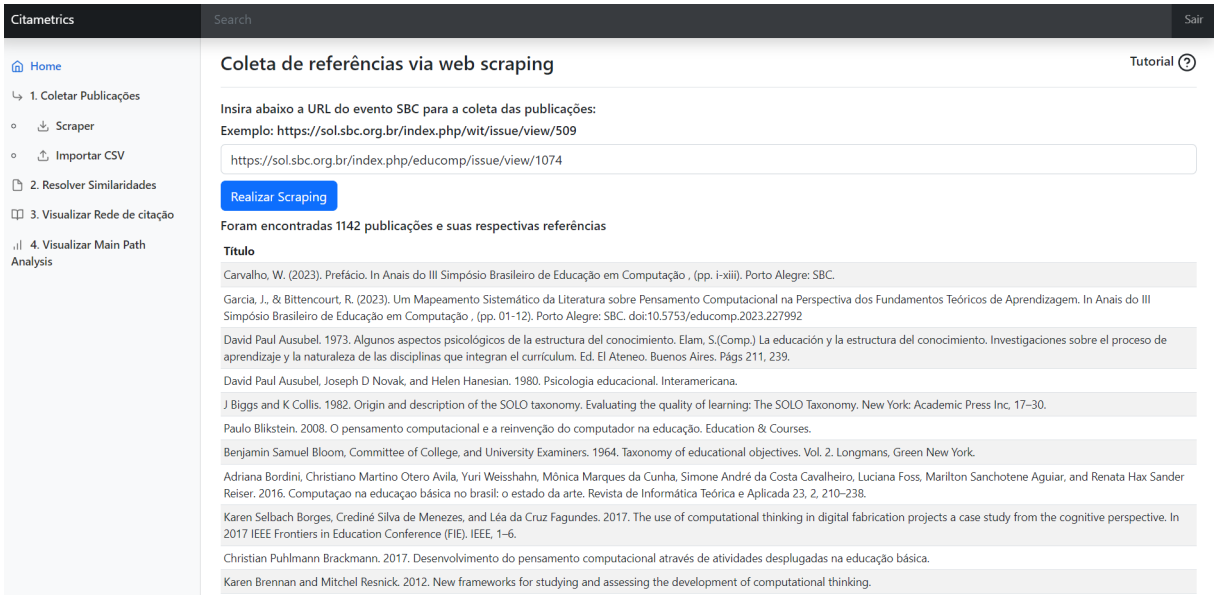


Fonte: Elaborado pelo autor (2023)

ser visto, as duas setas vermelhas da figura indicam pares de publicações similares onde ambas se referem a uma publicação única.

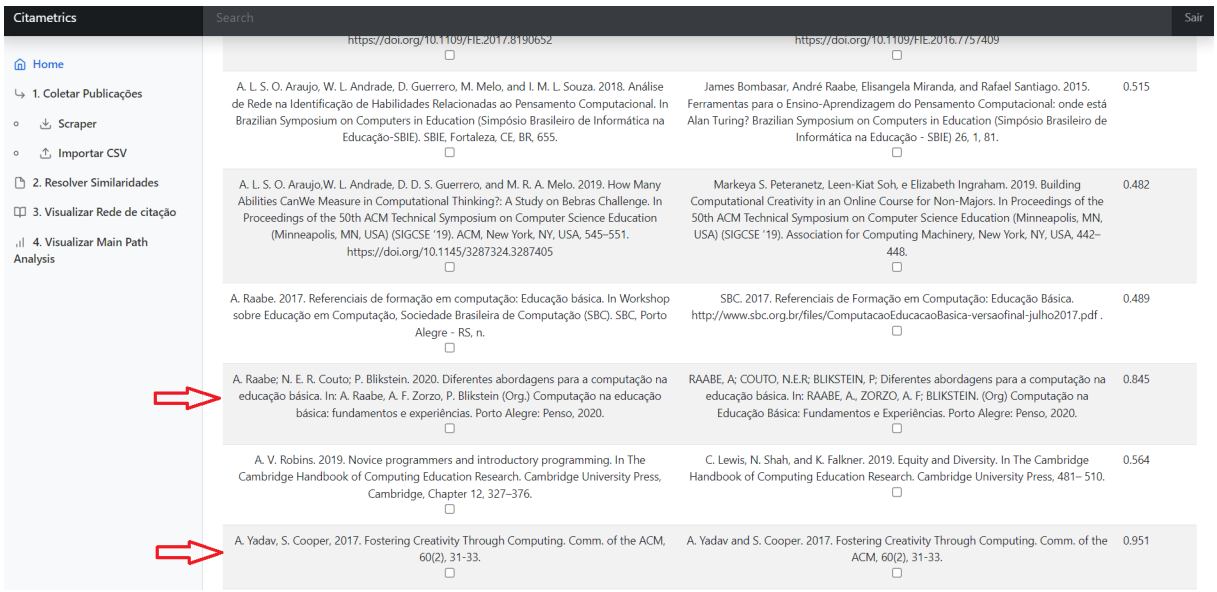
Essa funcionalidade é realizada por meio de uma tabela interativa, onde o usuário selecionada cada par de publicações similares, clicando nas linhas da tabela. É possível que existam múltiplas linhas com publicações similares, onde todas se referem a mesma publicação, como visto na Figura 28. Este comportamento é normal, e todas as linhas respectivas a mesma publicação devem ser selecionadas. O *checkbox* define que aquela

Figura 26 – Tabela com as publicações coletadas do ano de 2023



Fonte: Elaborado pelo autor (2023)

Figura 27 – Página para resolver similaridades



Fonte: Elaborado pelo autor (2023)

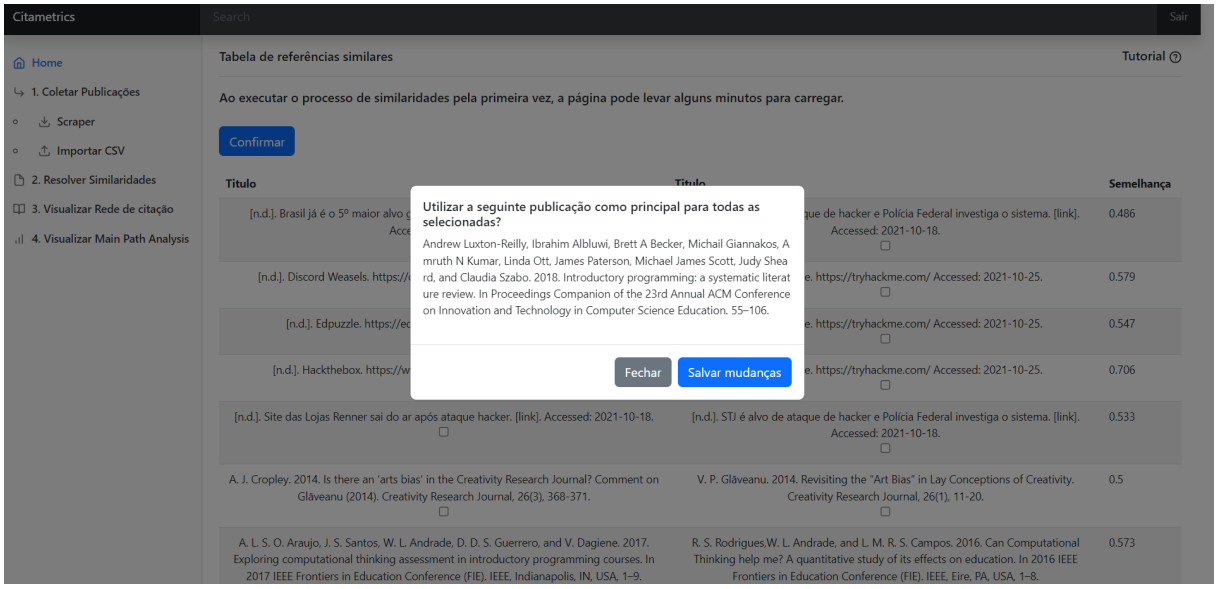
publicação será definida como a principal, substituindo todas as outras, portanto só é possível que um único *checkbox* esteja marcado. Novamente existe um tutorial para guiar o usuário nesta etapa crítica.

Há um botão para confirmar as escolhas do usuário, de acordo com a Figura 29. É importante que sejam selecionadas apenas publicações similares as que está informada como "utilizar como principal". Marcar publicações diferentes a esta fará com que a aplicação considere ambas como iguais, o que não é verdade e invalidará a análise subsequente.

Figura 28 – Tabela interativa

Fonte: Elaborado pelo autor (2023)

Figura 29 – Confirmação das mudanças



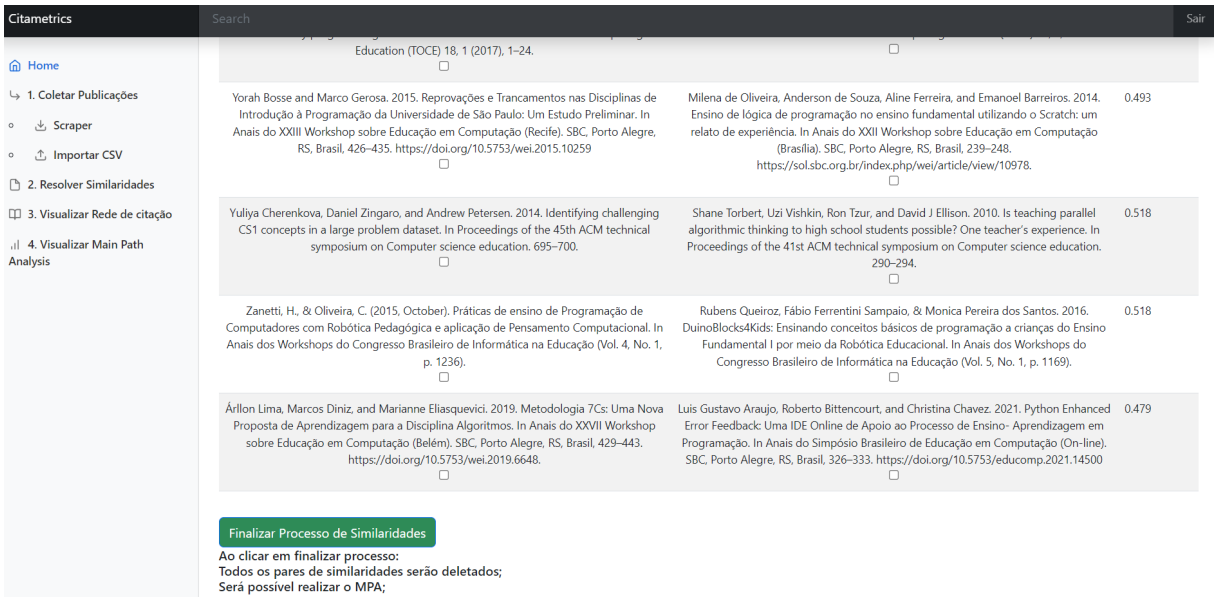
Fonte: Elaborado pelo autor (2023)

Ao clicar em salvar mudanças a página é recarregada imediatamente, refletindo as mudanças informadas pelo usuário, que já foram salvas no banco de dados. Isso permite que o progresso fique salvo mesmo que o usuário feche a página ou faça um *logout*. Ou seja, não é necessário que essa etapa seja completada de uma vez só, o que aumenta muito a conveniência do uso.

Quando o usuário julgar que não há mais publicações similares, o processo é encerrado ao clicar no botão finalizar similaridades, como na Figura 30. Existe um *pop-up*

para confirmar a finalização, prevenindo erros.

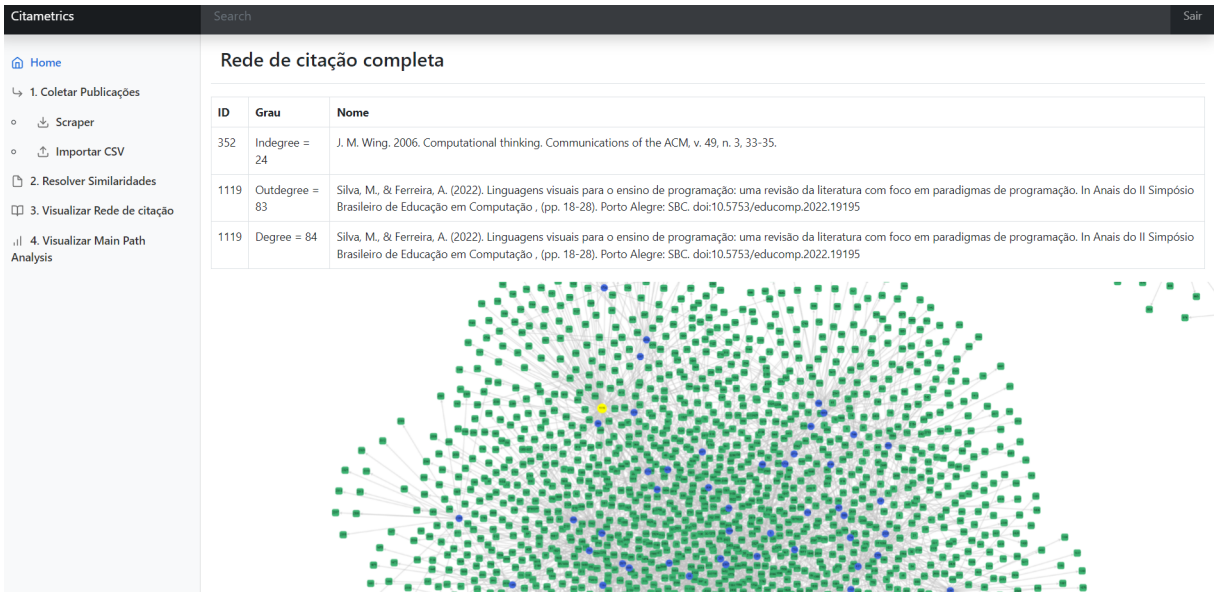
Figura 30 – Finalizar processo de similaridades



Fonte: Elaborado pelo autor (2023)

Ao finalizar o processo de similaridades, o usuário é levado a uma página com a rede de citações completa, bem como algumas informações pertinentes, vistas na Figura 31.

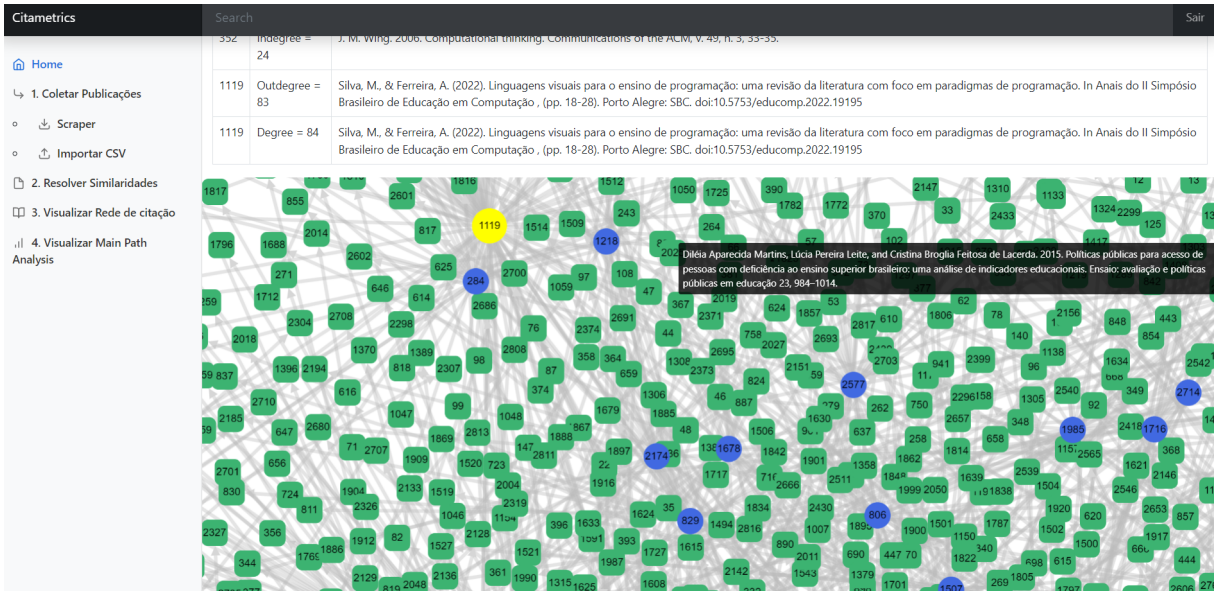
Figura 31 – Rede de citação completa



Fonte: Elaborado pelo autor (2023)

O gráfico é interativo, sendo possível mover os nós, alterar o zoom, e, quando o mouse passa por cima de um nó, o título da publicação é exibido, de acordo com a Figura 32.

Figura 32 – Grafo interativo



Fonte: Elaborado pelo autor (2023)

A última etapa do processo é a visualização da rede do MPA, como visto na Figura 33. Esta página possui dois gráficos, o primeiro é o gráfico principal, contendo apenas os vértices do caminho crítico. Novamente ao passar o mouse por cima dos nós o título completo da publicação é exibido.

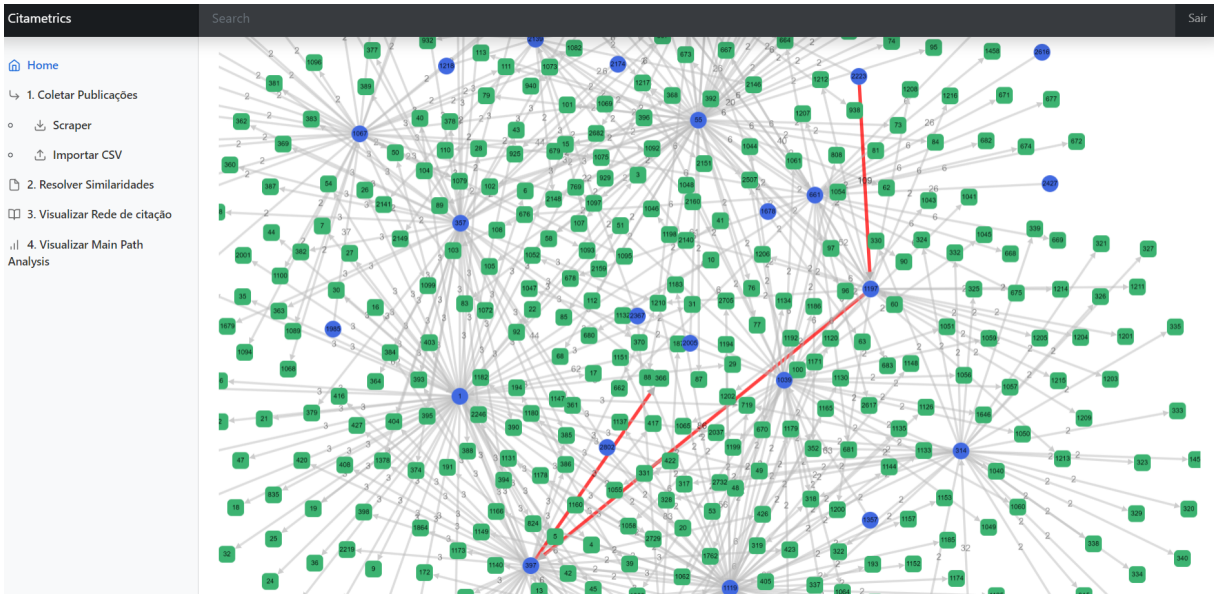
Figura 33 – Rede do caminho crítico

Fonte: Elaborado pelo autor (2023)

O outro grafo é a rede de citações completa, onde apenas as arestas com o resultado de cálculo do SPLC maior que 1 são desenhadas. Isto foi feito para melhorar a visualização

de grafos grandes, como na Figura 34. Por fim as arestas em vermelho destacam o caminho crítico, o mesmo que é exibido no primeiro grafo, mas aqui no grafo completo.

Figura 34 – Rede de citações com o MPA



Fonte: Elaborado pelo autor (2023)

Após a realização deste estudo, foi possível concluir que a ferramenta auxilia o usuário nas diversas etapas de uma análise de rede de citação, desde a coleta e o processamento dos dados, até a visualização dos resultados. Entretanto, existem alguns pontos de atenção, como por exemplo a elevada especificidade da mesma, requerendo um usuário muito especializado para tirar um bom proveito na aplicação. Também só um método de análise bibliométrica foi implementado, mas o futuro outros métodos podem ser incluídos na aplicação.

7 CONSIDERAÇÕES FINAIS

O objetivo deste trabalho foi desenvolver uma aplicação *web*, *open source*, que analisa redes de citação com o método bibliométrico *Main Path Analysis*. De acordo com o levantamento das ferramentas bibliométricas existentes, foram identificadas algumas lacunas não atendidas pelas ferramentas analisadas: uma delas diz respeito a inexistência de uma aplicação *web*, uma vez que todas as outras ferramentas existentes são aplicações para *desktop*; outra lacuna é o fato do método de *MPA* estar implementado em somente uma ferramenta, sendo a referida ferramenta notoriamente difícil de utilizar.

Os testes realizados e a utilização com sucesso da ferramenta **CitaMetrics** em um cenário de uso completo, evidenciam que os objetivos gerais e específicos foram cumpridos, uma vez que a aplicação permite realizar o fluxo completo de análise de uma rede de citação, desde a coleta dos dados via *web scraping*, até a execução do método de MPA, exibindo os resultados em um grafo interativo para o usuário.

Dessa maneira, a aplicação se mostra adequada para ser usada por pesquisadores interessados em análises baseadas em redes de citação com foco no método do MPA. A aplicação permite estudar o histórico do desenvolvimento de uma área, assim como suas principais publicações. Também facilita o processo de análise de citações, sendo a única ferramenta existente atualmente com funcionalidades para eliminação de redundâncias e normalização das referências bibliográficas. Esta, inclusive, foi uma das motivações originais do desenvolvimento da ferramenta, decorrente do trabalho hercúleo de coleta e análise manual de referências bibliográficas do evento *Women in Technology* realizado por Bordin *et al.* (2022). Além disso, sendo uma aplicação web, é mais fácil de utilizar, não requerendo instalação prévia. Por fim, é *open source* e está disponível em repositório público da UFSC, logo pode ser estendida pela comunidade.

Entretanto, a aplicação possui algumas limitações e, por consequência, oportunidades para trabalhos futuros. A coleta de dados automática via *web scraping* foi desenvolvida para uma única base de dados de publicações científicas, a SBC Open Library (SOL). Esta base foi escolhida pois indexa eventos e periódicos nacionais da grande área da Computação com um bom extrato Qualis, contendo inúmeras publicações. Além disso, todo o conteúdo da base de dados é de livre acesso, e a presença de dados das publicações no corpo das páginas HTML simplifica muito a coleta utilizando *Web Scraping*.

O desenvolvimento do *scraper* foi desafiador, devido a problemas como a falta de padronização da biblioteca SOL no que se refere à forma de apresentação dos dados de publicações e respectivas referências bibliográficas. No futuro podem ser implementados *scrapers* para outras bases de dados. Por fim, a funcionalidade de importar publicações somente é compatível com arquivos CSV, com uma estrutura interna específica. Sugere-se que seja incluído em versões futuras a opção de importar arquivos JSON, mais adequados a redes de citação.

O processo de identificação de similaridades, apesar deste ser substancialmente mais rápido e fácil do que uma revisão manual, pode ser um tanto moroso para o usuário, sofrendo com falsos positivos devido a grande variação de como as citações são escritas. Como trabalho futuro, sugere-se que sejam coletados dados de inúmeros eventos, valendo-se do *web scraper* já implementado, a fim de treinar uma Inteligência Artificial (IA) para resolver o processo de similaridades automaticamente.

Sugere-se também que sejam incluídas funcionalidades colaborativas. Atualmente, cada usuário tem uma lista única de projetos, mas no futuro isto poderia ser modificado para que vários usuários compartilhassem um projeto, dessa maneira permitindo a colaboração de vários indivíduos nas análises efetuadas. Isso permitiria não só uma maior conveniência aos usuários, facilitando o compartilhamento do trabalho, como também resultaria em análises mais robustas.

Além disso, até o momento foi desenvolvido um único método de análise bibliométrica, o MPA. Sugere-se, portanto, que sejam implementados outros métodos de análise de citação, como o acoplamento bibliográfico, a co-citação, assim como métricas tradicionais de análise de rede, como as métricas de centralidade. Apesar do foco da aplicação ter sido na análise de redes de citações, o MPA também pode ser utilizado na análise de patentes, e a partir de tal análise, realizar *technology forecasts*, portanto sugere-se a expansão da ferramenta para incluir tais funcionalidades.

Por fim, no quesito de usabilidade, não foi possível que outros usuários testassem a ferramenta, visto que essa é de um nicho muito específico e houve falta de tempo hábil. Dessa forma, sugere-se a seleção de usuários especialistas em análises de redes de citação para participação em testes de usabilidade e de aceitação de tecnologia, como o *Technology Acceptance Model (TAM).*

Este trabalho serviu como um grande aprendizado na área de programação web. Muitos obstáculos foram superados, tais como entender a comunicação assíncrona entre cliente e servidor, a utilização de bancos de dados *nosql*, a lógica para tratar da normalização das publicações, o desenvolvimento de telas interativas da aplicação, além da própria implementação do MPA.

REFERÊNCIAS

ARAÚJO, Carlos Alberto. Bibliometria: evolução histórica e questões atuais. **Em Questão**, 12), p. 11–32, 2006. ISSN 1807-8893. Disponível em: https://www.redalyc.org/articulo.oa?id=465645954002.

ARIA, Massimo; CUCCURULLO, Corrado. bibliometrix: An R-tool for comprehensive science mapping analysis. **Journal of informetrics**, Elsevier, v. 11, n. 4, p. 959–975, 2017.

BARABÁSI, Albert-László. Network science. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, The Royal Society Publishing, v. 371, n. 1987, p. 20120375, 2013.

BARBIERI, Nicolò *et al.* A survey of the literature on environmental innovation based on main path analysis. **Environmental Economics and Sustainability**, Wiley Online Library, p. 221–250, 2017.

BATAGELJ, Vladimir. Efficient algorithms for citation network analysis. **arXiv preprint cs/0309023**, 2003.

BATAGELJ, Vladimir; MRVAR, Andrej. Pajek-program for large network analysis. **Connections**, v. 21, n. 2, p. 47–57, 1998.

BORDIN, Andréa *et al.* Uma Análise das Citações do Women in Technology (WIT). *In:* ANAIS do XVI Women in Information Technology. Niterói: SBC, 2022. P. 157–166. DOI: 10.5753/wit.2022.223279. Disponível em: https://sol.sbc.org.br/index.php/wit/article/view/20868.

BORGATTI, Stephen P; EVERETT, Martin G; FREEMAN, Linton C. Ucinet for Windows: Software for social network analysis. **Harvard, MA: analytic technologies**, v. 6, p. 12–15, 2002.

BOYACK, Kevin W; KLAVANS, Richard. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? **Journal of the American Society for information Science and Technology**, Wiley Online Library, v. 61, n. 12, p. 2389–2404, 2010.

CARVALHO, Maria Martha de. Análises bibliométricas da literatura de química no Brasil. **Ciência da Informação**, v. 4, n. 2, dez. 1975. DOI: 10.18225/ci.inf.v4i2.56. Disponível em: https://revista.ibict.br/ciinf/article/view/56.

CASTANHA, Rafael Gutierres; SANTOS JÚNIOR, Edmilson Alves dos; TOLARE, Jéssica Beatriz. Cultura da convergência: uma análise a partir dos indicadores bibliométricos de produção, citação e relacional de cocitação de autores na base de dados Web of Science (2008-2021). **Em Questão**, SciELO Brasil, v. 29, e–122198, 2023.

CHEN, Chaomei. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. **Journal of the American Society for information Science and Technology**, Wiley Online Library, v. 57, n. 3, p. 359–377, 2006.

CHEN, Liang *et al.* A semantic main path analysis method to identify multiple developmental trajectories. **Journal of Informetrics**, v. 16, n. 2, p. 101281, 2022. ISSN 1751-1577. DOI: https://doi.org/10.1016/j.joi.2022.101281. Disponível em: https://www.sciencedirect.com/science/article/pii/S1751157722000335.

CORMEN, Thomas H. Single-source shortest paths. **Introduction to algorithms**, MIT press, 2001.

DAMERAU, Fred J. A technique for computer detection and correction of spelling errors. **Communications of the ACM**, ACM New York, NY, USA, v. 7, n. 3, p. 171–176, 1964.

DAWSON, Shane *et al.* Current state and future trends: A citation network analysis of the learning analytics field. *In:* PROCEEDINGS of the fourth international conference on learning analytics and knowledge. [*S.l.*: *s.n.*], 2014. P. 231–240.

DEMETRESCU, Camil *et al.* On bibliometrics in academic promotions: a case study in computer science and engineering in Italy. **Scientometrics**, Springer, v. 124, n. 3, p. 2207–2228, 2020.

DONTHU, Naveen *et al.* How to conduct a bibliometric analysis: An overview and guidelines. **Journal of business research**, Elsevier, v. 133, p. 285–296, 2021.

EVEN, Shimon. **Graph algorithms**. [*S.l.*]: Cambridge University Press, 2011.

FU, Hanliang *et al.* Tracing knowledge development trajectories of the internet of things domain: a main path analysis. **IEEE Transactions on Industrial Informatics**, IEEE, v. 15, n. 12, p. 6531–6540, 2019.

GARFIELD, Eugene; MERTON, Robert K. **Citation indexing: Its theory and application in science, technology, and humanities**. [*S.l.*]: Wiley New York, 1979. v. 8.

GRÁCIO, Maria Cláudia Cabrini. A coplamento bibliográfico e análise de cocitação: revisão teórico-conceitual. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 21, n. 47, p. 82–99, 2016.

GRÁCIO, Maria Claudia Cabrini. **Análises relacionais de citação para a identificação de domínios científicos: uma aplicação no campo dos Estudos Métricos da Informação no Brasil**. [*S.l.*]: Editora UNESP, 2020. DOI: 10.36311/2020.978-65-86546-12-5.

HJØRLAND, Birger. User-based and cognitive approaches to knowledge organization: A theoretical analysis of the research literature. **KO KNOWLEDGE ORGANIZATION**, Nomos Verlagsgesellschaft mbH & Co. KG, v. 40, n. 1, p. 11–27, 2014.

HOTA, Pradeep Kumar; SUBRAMANIAN, Balaji; NARAYANAMURTHY, Gopalakrishnan. Mapping the intellectual structure of social entrepreneurship research: A citation/co-citation analysis. **Journal of business ethics**, Springer, v. 166, n. 1, p. 89–114, 2020.

HUMMON, Norman P; DEREIAN, Patrick. Connectivity in a citation network: The development of DNA theory. **Social networks**, Elsevier, v. 11, n. 1, p. 39–63, 1989.

JARNEVING, Bo. Bibliographic coupling and its application to research-front and other core documents. **Journal of Informetrics**, v. 1, n. 4, p. 287–307, 2007. ISSN 1751-1577. DOI: https://doi.org/10.1016/j.joi.2007.07.004. Disponível em: https://www.sciencedirect.com/science/article/pii/S1751157707000594.

KESSLER, Maxwell Mirton. Bibliographic coupling between scientific papers. **American documentation**, Wiley Online Library, v. 14, n. 1, p. 10–25, 1963.

KORTE, Andreas; TIBERIUS, Victor; BREM, Alexander. Internet of Things (IoT) Technology Research in Business and Management Literature: Results from a Co-Citation Analysis. **Journal of Theoretical and Applied Electronic Commerce Research**, v. 16, n. 6, p. 2073–2090, 2021. ISSN 0718-1876. DOI: 10.3390/jtaer16060116. Disponível em: https://www.mdpi.com/0718-1876/16/6/116.

LIU, John S; LU, Louis YY. An integrated approach for main path analysis: Development of the Hirsch index as an example. **Journal of the American Society for Information Science and Technology**, Wiley Online Library, v. 63, n. 3, p. 528–542, 2012.

LIU, John S; LU, Louis YY; HO, Mei Hsiu-Ching. A few notes on main path analysis. **Scientometrics**, Springer, v. 119, p. 379–391, 2019.

LIU, John S.; LU, Louis Y.Y.; LU, Wen-Min *et al.* Data envelopment analysis 1978–2010: A citation-based literature survey. **Omega**, v. 41, n. 1, p. 3–15, 2013. Data Envelopment Analysis: The Research Frontier - This Special Issue is dedicated to the memory of William W. Cooper 1914-2012. ISSN 0305-0483. DOI: https://doi.org/10.1016/j.omega.2010.12.006. Disponível em: https://www.sciencedirect.com/science/article/pii/S0305048312000291.

LUCAS, Elaine Oliveira; GARCIA-ZORITA, Jose Carlos. Produção científica sobre capital social: estudo por acoplamento bibliográfico. **Em Questão**, Universidade Federal do Rio Grande do Sul, v. 20, n. 3, p. 27–42, 2014.

MA, Tsu-Jui *et al.* Bibliographic coupling: A main path analysis from 1963 to 2020. University of Borås, 2022.

MARSHAKOVA, Irina. Citation networks in information science. **Scientometrics**, Akadémiai Kiadó, co-published with Springer Science+ Business Media BV ..., v. 3, n. 1, p. 13–25, 1981.

MASEDA, Amaia *et al.* Mapping women's involvement in family firms: A review based on bibliographic coupling analysis. **International Journal of Management Reviews**, Wiley Online Library, v. 24, n. 2, p. 279–305, 2022.

PERSSON, Olle; DANELL, Rickard; SCHNEIDER, J Wiborg. How to use Bibexcel for various types of bibliometric analysis. **Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th Birthday**, v. 5, p. 9–24, 2009.

PHAN TAN, Luc. Bibliometrics of social entrepreneurship research: Cocitation and bibliographic coupling analyses. **Cogent Business & Management**, Taylor & Francis, v. 9, n. 1, p. 2124594, 2022.

PRICE, Derek J De Solla. Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. **Science**, American Association for the Advancement of Science, v. 149, n. 3683, p. 510–515, 1965.

RAFAEL, Diego Nogueira; HERRERO, Eliane; SOUSA, Eduardo Mesquita de. ESTRUTURA INTELECTUAL DO EFEITO PLACEBO: ANÁLISE DE COCITAÇÃO DAS ÚLTIMAS DUAS DÉCADAS. **Revista Pretexto**, v. 24, n. 2, 2023.

SMALL, Henry. Co-citation in the scientific literature: A new measure of the relationship between two documents. **Journal of the American Society for information Science**, Wiley Online Library, v. 24, n. 4, p. 265–269, 1973.

VAN ECK, Nees; WALTMAN, Ludo. Software survey: VOSviewer, a computer program for bibliometric mapping. **scientometrics**, Akadémiai Kiadó, co-published with Springer Science+ Business Media BV ..., v. 84, n. 2, p. 523–538, 2010.

WHITE, Howard D; GRIFFITH, Belver C. Author cocitation: A literature measure of intellectual structure. **Journal of the American Society for information Science**, Wiley Online Library, v. 32, n. 3, p. 163–171, 1981.

YAN, Jianghui; TSENG, Fang-Mei; LU, Louis Y.Y. Developmental trajectories of new energy vehicle research in economic management: Main path analysis. **Technological Forecasting and Social Change**, v. 137, p. 168–181, 2018. ISSN 0040-1625. DOI: `https://doi.org/10.1016/j.techfore.2018.07.040`. Disponível em: `https://www.sciencedirect.com/science/article/pii/S0040162517312064`.

YU, Dejian; SHENG, Libo. Knowledge diffusion paths of blockchain domain: the main path analysis. **Scientometrics**, Springer, v. 125, n. 1, p. 471–497, 2020.

ZHANG, Ben; MA, Lei; LIU, Zheng. Literature Trend Identification of Sustainable Technology Innovation: A Bibliometric Study Based on Co-Citation and Main Path Analysis. **Sustainability**, v. 12, n. 20, 2020. ISSN 2071-1050. DOI: 10.3390/su12208664. Disponível em: https://www.mdpi.com/2071-1050/12/20/8664.

ZHAO, Dangzhi; STROTMANN, Andreas. Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic-coupling analysis. **Journal of the American Society for Information Science and Technology**, Wiley Online Library, v. 59, n. 13, p. 2070–2086, 2008.

APÊNDICE A – DOCUMENTAÇÃO

Neste apêndice é apresentada uma breve documentação das funcionalidades Realizar Web Scraping e Resolver Referências Similares. Também, o código fonte da aplicação Citametrics se encontra no repositório de códigos *Gitlab* da UFSC, disponível em: https://codigos.ufsc.br/ramom.giovani.guth/mpa_analysis

Realizar Web Scraping: Por meio desta funcionalidade são coletadas, via web scraping, as referências bibliográficas de trabalhos citados por trabalhos publicados em anais de eventos e periódicos. Esse é um processo totalmente automatizado que coleta e salva as referências. O único input do usuário é o link do evento/periódico a ser analisado. Atualmente o web scraper é compatível com a base de dados da SBC (SOL).

Quando o usuário insere a URL de um evento/periódico são coletadas as URLs de todas as publicações apresentadas na URL inserida. A partir de cada URL de publicação são extraídas e armazenadas informações como a referência da própria publicação, bem como sua a lista das referências de trabalhos que a publicação citou.

Esse processo retorna para o usuário, no formato de uma tabela, todas as referências de publicações indexadas em uma determinada edição de evento/periódico e as respectivas referências de trabalhos citados pela publicações. O processo pode ser repetido quantas vezes o usuário desejar, bastando alterar o link do evento/periódico a ser analisado. Existe um controle para prevenir a inclusão de links/eventos repetidos.

Resolver Referências Similares: Coletadas as referências bibliográficas dos trabalhos, se faz necessário eliminar referências redundantes. Cabe ressaltar que um mesmo trabalho pode ser citado várias vezes, o que faz com que referências a este trabalho sejam extraídas várias vezes. Além disso, um trabalho pode ser citado/referenciado por outros trabalhos de maneiras diferentes. A referência a um trabalho deve ser representada por meio de um nó/nodo uma única vez em uma rede/grafo de citação.

Para resolver esse problema, a aplicação auxilia o usuário comparando uma referência com todas as outras, par a par,fazendo efetivamente uma análise combinatória, dando uma nota de o quanto uma referência é similar a outra. Essa comparação é utilizada a distância de *Damerau-Levenshtein*. Caso a nota da distância seja superior a um *threshold* pré-definido, a aplicação irá salvar essa esse par de publicações similares, bem como a nota de similaridade, no banco de dados. Posteriormente será exibido para o usuário em uma tabela os pares de publicações similares. Para fins de exemplificação, esse processo levou aproximadamente 2 segundos para fazer a comparação de 550 publicações.

Tal tabela conterá inúmeros pares de referências/publicações similares, e caberá então ao usuário analisar os pares similares, indicando quais pares se referem à mesma publicação. Caso um ou mais pares de referências mostradas se refiram à mesma publicação, o usuário escolherá uma única publicação que servirá como principal, substituindo todas as outras.

Essa análise tem a capacidade de ser realizada em partes, isto é, o usuário poderá fazer logout e continuar a análise em um outro momento do mesmo ponto de onde parou. Quando o usuário terminar o processo de resolução de similaridades, clicando em um botão específico para esta ação, a aplicação irá permitir que enfim seja calculado o MPA.