Felipe Ciacia de Mendonça

Geração das Redes de Colaboração Científica da Comunidade Acadêmica de IHC

Joinville

2017

UNIVERSIDADE DO ESTADO DE SANTA CATARINA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Felipe Ciacia de Mendonça

GERAÇÃO DAS REDES DE COLABORAÇÃO CIENTÍFICA DA COMUNIDADE ACADÊMICA DE IHC

Trabalho de conclusão de curso submetido à Universidade do Estado de Santa Catarina como parte dos requisitos para a obtenção do grau de Bacharel em Ciência da Computação

Joinville, Novembro de 2017

GERAÇÃO DAS REDES DE COLABORAÇÃO CIENTÍFICA DA COMUNIDADE ACADÊMICA DE IHC

Felipe Ciacia de Mendonça

Este Trabalho de Conclusão de Curso foi julgado adequado para a obtenção do título de Bacharel em Ciência da Computação e aprovado em sua forma final pelo Curso de Ciência da Computação Integral do CCT/UDESC.

Banca Examinadora

Isabela Gasparini - Doutora - UDESC
(orientador)

Rebeca Schroeder Freitas - Doutora - UDESC

Milene Selbach Silveira - Doutora - PUCRS

Simone Diniz Junqueira Barbosa - Doutora - PUC-Rio

Agradecimentos

A Deus que me deu força, saúde e determinação para superar as dificuldades encontradas no caminho e permitir que tudo isso acontecesse.

Á minha orientadora Professora Doutora Isabela Gasparini, minha eterna gratidão pela oportunidade de ser seu orientado tanto de bolsista de iniciação científica, quanto de trabalho de conclusão de curso. E também pelos seus sábios ensinamentos e dedicação dada à mim.

À Professora Doutora Rebeca Schroeder Freitas que aceitou fazer parte da banca, e também por ter me ajudado com seus ensinamentos e dicas durante o curso desde quando pude ter o prazer de ter aula em sua matéria.

Agradeço às professoras desta banca Milene Selbach Silveira e Simone Diniz Junqueira Barbosa por aceitarem fazer parte da banca, e por contribuirem na realização do artigo enviando ao Simpósio do IHC.

Agradeço a todos os professores por me proporcionarem o conhecimento não apenas técnico, mas a construção de caráter no processo de formação profissional.

Aos meus pais, pelo amor, incentivo, e apoio incondicional não só durante todos os anos de curso, mas durante toda a vida, principalmente nas horas mais difíceis.

A todos meus amigos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado!

"É muito melhor lançar-se em busca de conquistas grandiosas, mesmo expondo-se ao fracasso, do que alinhar-se com os pobres de espírito, que nem gozam muito nem sofrem muito, porque vivem numa penumbra cinzenta, onde não conhecem nem vitória, nem derrota." (Theodore Roosevelt)

Resumo

As redes sociais possibilitam a formação de grupos e comunidades específicas tais como as redes de colaboração científica, que são compostas por seus autores, e conectadas por meio de suas publicações. Este trabalho tem como objetivo gerar e analisar as redes de colaboração científica da Comunidade Brasileira de Interação Humano-Computador (IHC) analisando os autores mais prolíficos do Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (Simpósio IHC). Por meio da ferramenta de extração desenvolvida neste trabalho, foi realizada a extração dos dados dos autores mais prolíficos registrados na Plataforma Lattes, além da realização do entendimento, preparação e modelagem dos dados com base nas técnicas de classificação, limpeza e padronização de mineração de dados. Por fim, foram geradas e analisadas as redes de colaboração científica por meio de análises de redes sociais e análises bibliométricas. Assim, foram extraídos resultados sobre a comunidade brasileira de IHC em relação a padrões, relacionamentos, *insights*, além da própria contribuição da ferramenta implementada.

Palavras-chaves: IHC, Redes de Colaboração Científica, Análise de Redes Sociais, Análises Bibliométricas, Mineração de Dados

Abstract

Social networks enable the formation of specific groups and communities such as scientific collaboration networks, which are composed of their authors, and connected through their publications. This work aims to generate the networks of scientific collaboration of the Brazilian Community of Human-Computer Interaction (IHC) analyzing the most prolific authors of the Brazilian Symposium on Human Factors in Computing Systems (Simpósio IHC). Through the extraction tool developed in this work, we extracted the data from the most prolific authors registered in the Lattes Platform, in addition to the understanding, preparation and modeling of the data based on the techniques of classification, cleaning and standardization of data mining. Finally, scientific collaboration networks were generated and analyzed through the bibliometric analysis and techniques of social network analysis. Thus, results were extracted on a Brazilian community of IHC in relation to patterns, relationships, insights, besides the own contribution of the implemented tool.

Keywords: CHI, HCI, Scientific Collaboration Networks, Social Network Analysis, Data Mining, Bibliometric Analysis.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AA *Adamic-Adar*

ARS Análise de Redes Sociais

ASP *Active Server Pages*

BDB *Berkeley Data Base*

CAPES Comissão de Aperfeiçoamento de Pessoal do Nível Superior

CN *Common Neighbors*

CRISP-DM *Cross Industry Standard Process of Data Mining*

CSV *Comma-separated values*

Cypher Linguagem de Pesquisas em Grafos utilizada pelo Neo4J

DTD *Document Type Definition*

ENANCIB Encontro Nacional de Pesquisa em Ciência da Informação

ER Entidade-Relacionamento

FID Federação Internacional de Documentação

GIS *Geographic Information System*

GPL *General Public License*

GranDIHC-BR Grandes Desafios de Pesquisa em Interação Humano-Computador no Brasil.

GT2 Grupo de Trabalho "Organização e representação do conhecimento"@

HTML *HyperText Markup Language*

IHC Interação Humano-Computador

ISV *Independent Software Vendor*

JC *Jaccard Coefficient*

JSON *JavaScript Object Notation*

KDD *Knowledge Discovery in Database*

MyISAM Antigo mecanismo de armazenamento do *MySQL* para versões anteriores

MySQL Sistema de Gerenciamento de Banco de Dados

Neo4J Banco de Dados Orientado a Grafos

NiCHE *Network in Canadian History Environment*

NoSQL Banco de Dados Não Relacional

OEM *Original Equipment Manufacturer*

OLAP *Online Analytical Processing*

PA *Preferencial Attachment*

PHP *Personal Home Page*

PPG Programa de Pós Graduação

SBC Sociedade Brasileira de Computação

SimposioIHC Simposio Brasileiro sobre Fatores Humanos em Sistemas Computacionais

SGBD Sistema de Gerenciamento de Banco de Dados

SGML *Standard Generalized Markup Language*

SNPG Sistema Nacional de Pós-Graduação

SQL *Structured Query Language*

SUCUPIRA Sistema Unificado de Currículos e Programas: Identificação de Redes Acadêmicas

UDESC Universidade do Estado de Santa Catarina

URL *Uniform Resource Locator*

W3C *World Wide Web Consortium*

XML *eXtensible Markup Language*

YAML *Yet Another Multicolumn Layout*

1 Introdução

Diversos estudos sociológicos apontam que existe uma tendência natural para pessoas com características em comum se agruparem e formarem comunidades - famílias, amigos e grupos com afinidades (GABARDO, 2015). Essas comunidades se agrupam de acordo com as similaridades que os membros compartilham, onde as ligações entre os membros de uma rede sugerem um relacionamento entre eles (NEWMAN, 2001).

O crescimento da Web permitiu a criação de diversas ferramentas para a interação das comunidades, tais como as redes sociais online (por exemplo, Facebook, Twitter, LinkedIn, etc.), possibilitando a conexão entre as pessoas, a formação de grupos e comunidades específicas e o compartilhamento de conteúdo (BENEVENUTO; ALMEIDA; SILVA, 2011). As redes sociais online estão cada vez mais presentes atualmente, e são uma abstração da (real) rede social existente e podem ser interpretadas como interações entre as pessoas no mundo real (VISWANATH et al., 2009).

Devido à popularidade das redes sociais, e consequentemente a enorme quantidade de dados produzidos, a análise das redes sociais tem atraído atenção da comunidade científica, de forma a melhor compreender o comportamento das interações humanas (GABARDO, 2015). Essas ligações entre as pessoas em uma rede social podem ter diversos significados, como, amizades, família, profissional, entre outros (MARUYAMA; DIGIAMPETRI, 2016)

Entre as diversas redes sociais existentes, tem-se as redes formadas no âmbito científico que são as redes de colaboração científica ou redes acadêmicas. Nestas redes, os vértices (nós) representam os pesquisadores, e as arestas (ligações) as colaborações científicas. Portanto, se dois pesquisadores estão conectados, então eles são coautores em uma ou mais publicações (NEWMAN, 2001). Neste contexto, pode-se extrair explicitamente uma rede de colaboração científica direta entre os autores por meio de publicações de um ou mais veículos de comunicação. Ainda pode existir a rede de colaboração científica formada pela rede de citações, a qual seria uma forma de colaboração indireta, porém, neste trabalho focaremos apenas na rede de coautoria que é uma forma direta de colaboração científica.

Segundo Weisz e Roco (1996), a colaboração científica pode ser um empreendimento cooperativo que procura o desenvolvimento de trabalhos envolvendo metas em comum, esforço coordenado e resultados ou produtos (trabalhos científicos) com responsabilidade e mérito compartilhados. Assim, os autores consideram que a colaboração científica oferece uma fonte de apoio para melhorar o resultado e maximizar o potencial da produção científica.

Porém, apesar da colaboração científica existir há muito tempo, a análise e investigação das redes de coautoria para explorar a colaboração científica entre pesquisadores constituem uma área relativamente nova (UDDIN et al., 2012). O processo de disseminação de informações em redes sociais não é simples. A assimilação das informações pelos membros da rede depende de uma série de fatores, como por exemplo, interesse, percepção acerca do contexto, confiabilidade, dentre outros fatores sociológicos, psicológicos e antropológicos (GABARDO, 2015).

Segundo Gabardo (2015), compreender quais são os usuários mais influentes em uma rede social científica é de grande interesse a todos os pesquisadores. Assim, pode-se dizer que os membros com o maior número de conexões são os pontos mais valiosos para espalhar uma notícia ou informação. Também é possível saber quais são os autores mais produtivos e influentes dentro de uma comunidade, detectar os membros quanto à sua região geográfica, às suas produções, a internacionalização de sua pesquisa, às instituições nas quais é vinculado, e as próprias redes de coautoria das quais faz parte. Compreender como se formam essas redes de colaboração científica é importante para entender as preferências, padrões, e como se dá esse relacionamento entre os autores de uma comunidade científica (SILVA; BARBOSA; DUARTE, 2012)

A análise de redes sociais está ligada a conceitos aplicados à teoria dos grafos, e desta forma, grafos podem ser utilizados para representar as relações entre os autores. A análise de redes sociais também envolve algumas técnicas de mineração de dados que é o processo de descoberta de informações úteis em grandes repositórios de dados (TAN; STEINBACH; KUMAR, 2009). Conforme Tan, Steinbach e Kumar (2009) mencionam, essas técnicas também podem ser usadas para realizar tarefas de previsão, que têm como objetivo prever uma informação baseada em informações passadas.

Apoiado pelo trabalho de Gasparini et al. (2014), que analisou as redes de coautoria do próprio Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computa-

cionais (Simpósio IHC), este trabalho extraiu os autores mais prolíficos da comunidade brasileira de IHC por meio da base de dados do Simpósio (dados já obtidos por Gasparini et al. (2014)), mas estendeu a pesquisa, visto que analisou a rede de colaboração científica formada pelos membros mais prolíficos da comunidade brasileira de IHC. Para tal, foram observadas todas as publicações cadastradas na Plataforma Lattes destes autores nos diversos veículos de publicação.

Portanto, este trabalho visa gerar as redes de colaboração científica da comunidade brasileira de IHC, e para isso foi criado um modelo para transformação dos dados em informação (conhecimento) por meio, por exemplo, das técnicas de mineração de dados passando por diversas fases, entre elas a Pesquisa e Entendimento dos Dados, a Captura dos Dados, a Preparação dos Dados, e por fim, a Modelagem dos Dados. Os dados foram inicialmente capturados dos Anais do Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais selecionando os autores mais prolíficos que posteriormente, foram pesquisados na Plataforma Lattes para obtenção dos seus currículos registrados via formato XML (*eXtensible Markup Language*). A partir disso, foi implementada uma ferramenta em PHP (*PHP: Hypertext Preprocessor*) para extração automática dos dados dos currículos selecionados e assim, realizou-se a inserção em um banco de dados MySQL (Sistema de Gerenciamento de Banco de Dados) utilizando a linguagem SQL (*Structured Query Language*) para que fosse possível realizar o entendimento dos dados de forma mais clara, além de modelar e agrupar os mesmos para finalmente inserir em um banco de dados não relacional (*NoSQL*), que no caso foi utilizado o banco de dados orientado a grafos (*Neo4J*). Após a extração dos dados, foi realizada a padronização e limpeza das informações através das tabelas extraídas do banco de dados MySQL para que fosse inserida no banco de dados orientado a grafos. E por fim, no próprio banco de dados *Neo4J* foram geradas as redes de colaboração científica e realizadas as análises das redes de colaboração científica e dos dados extraídos da comunidade brasileira de IHC.

1.1 Objetivos

Esta seção apresenta o objetivo geral do trabalho e os objetivos específicos que foram alcançados para que o objetivo geral fosse realizado.

1.1.1 Objetivo Geral

Analisar as Redes de Colaboração Científica da Comunidade Brasileira de IHC extraindo os autores mais prolíficos do Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais e utilizando suas publicações nos diversos veículos de comunicação.

1.1.2 Objetivos Específicos

1. Extração dos dados dos Anais do Simpósio IHC e da Plataforma Lattes;
2. Implementação da ferramenta em PHP para extração automática dos dados dos currículos e inserção em uma nova base de dados;
3. Gerar as redes de colaboração científica da comunidade brasileira de IHC no bando de dados orientado a grafos, através das análises das redes sociais e de mineração de dados.

1.2 Metodologia

Este trabalho de conclusão de curso possui natureza aplicada, e é caracterizado como um trabalho de caráter experimental, pois foi necessário a aplicação de técnicas que foram analisadas no decorrer da pesquisa. A abordagem é na forma quantitativa já que foi feita a definição dos dados que foram utilizados para que então fosse traçada a configuração da rede e a apresentação das medidas que irão especificar os padrões de relacionamento entre os membros da rede. A pesquisa foi realizada começando com um levantamento bibliográfico sobre as técnicas de Análise de Redes Sociais e Complexas, assim como técnicas de Mineração de Dados, procurando a melhor forma de analisar os dados, que foram extraídos dos anais do IHC e dos currículos dos autores pesquisados e registrados na Plataforma Lattes, além de utilizar também trabalhos relacionados como base para realizar esse processo. Após ter realizado a fundamentação teórica, foram realizadas a captura dos dados, e posteriormente, a análise dos dados obtidos e a geração das redes de colaboração científica. Para atingir os objetivos propostos, foram realizadas as seguintes etapas:

1. Investigação de trabalhos relacionados

2. Seleção dos autores mais prolíficos de IHC através dos Anais do Simpósio IHC

3. Captura dos currículos selecionados através dos dados do Simpósio (IHC) e da Plataforma Lattes.

4. Implementação da ferramenta em PHP para extração automática de dados dos currículos em formato XML dos autores selecionados

5. Estudo, entendimento, preparação e modelagem dos dados para, avaliar e implantar as técnicas de análise de redes sociais.

6. Gerar e classificar as redes de colaboração científica da comunidade brasileira de IHC através do banco de dados orientado a grafos.

7. Análise dos Resultados

8. Escrita do trabalho de conclusão de curso

1.3 Estrutura do Trabalho

De acordo com o objetivo deste estudo, o trabalho foi estruturado em três capítulos. O Capítulo 2 descreve os fundamentos envolvidos neste trabalho, apresentando os conceitos de alguns termos que foram usados neste trabalho como a Rede de Colaboração Científica por exemplo, e também quais as técnicas que foram aplicadas. O Capítulo 3 apresenta os trabalhos que estão relacionados ao tema desta pesquisa, e que auxiliaram de alguma forma para utilizar como base para esse estudo. No Capítulo 4 é descrito a proposta deste trabalho mostrando todas as etapas que foram realizadas durante este trabalho. No Capítulo 5 são apresentadas as conclusões. Por fim, são apresentadas as referências bibliográficas utilizadas e os apêndices.

2 Fundamentação

Vários dos principais fundamentos que estão envolvidos no presente trabalho estão ligados a análises bibliométricas, análise de redes sociais e mineração de dados, e que são descritos nesta seção. Para facilitar a compreensão deste trabalho e dos conceitos envolvidos, na Seção 2.2 é destacado a pesquisa envolvendo a Comunidade Brasileira de IHC, já na Seção 2.2, são abordados os conceitos envolvendo a Bibliometria, Cienciometria e Infometria, e na Seção 2.3 são abordados os fundamentos da Análise de Redes Sociais. Na Seção 2.4 são apresentadas a Mineração de Dados e suas técnicas e por fim, na Seção 2.5 são apresentadas as considerações finais sobre o capítulo.

2.1 Comunidade Brasileira de IHC

A área de Interação Humano-Computador (IHC) é um campo de pesquisa que estuda como as pessoas interagem com os sistemas computacionais e até que ponto os computadores são ou não desenvolvidos para uma interação bem sucedida com os seres humanos (JONES, 2016). Essa área tem como um de seus objetivos investigar e produzir alternativas tecnológicas envolvidas no design e na avaliação de interfaces de usuários, para as pessoas interagirem de forma produtiva com métodos, técnicas, modelos e representações utilizados em diversos sistemas, tais como os sistemas Web, sistemas multimídia, sistemas em automóveis, smartphones, TVs digitais e nos próprios computadores tradicionais (SOUZA, 2006; BARBOSA; SILVA, 2010).

Uma forma de se conhecer quem trabalha com IHC, e quais trabalhos estão sendo produzidos é analisando os periódicos e conferências da área, e dentro desse contexto está o Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (Simpósio IHC) no qual constitui-se do principal encontro da comunidade brasileira de IHC composta por pesquisadores, professores e profissionais da área. Existem diversos trabalhos envolvendo a comunidade brasileira de IHC com o objetivo de estudar e compreender a comunidade por meio das publicações nos eventos a elas relacionados, porém, num primeiro momento apenas com o foco na comunidade brasileira através do Simpósio

IHC.

Entre os principais trabalhos com este propósito, destacam-se os trabalhos realizados por Gasparini e colegas (2013, 2014, 2015, 2016, 2017) e o trabalho realizado por Bueno et al. (2016). Nesses trabalhos foram realizadas análises bibliométricas, como número de artigos publicados por ano, principais autores, repetição/frequência de autores, questões de gênero, instituições e redes de coautoria, entre outros.

Gasparini, Kimura e Pimenta (2013) apresentam uma exploração visual de todas as 15 edições do IHC até o ano de 2013, permitindo uma visualização e descrição das informações coletadas a partir do conjunto de dados formado pelos artigos completos da conferência. Esta pesquisa em particular traz uma análise geral sobre diversas características dos dados coletados. Já em 2014, foi realizada a análise das redes de coautoria do IHC onde houve a investigação da colaboração dos autores de IHC analisando assim as redes de coautoria científica (GASPARINI et al., 2014).

No estudo de Gasparini, Silveira e Barbosa (2015), as autoras procuraram descobrir como tem ocorrido a migração dos pesquisadores de IHC no Brasil, além de quais seriam os principais centros de formação dos autores de artigos publicados na conferência. Outra análise interessante no mesmo ano de 2015 foi identificar como a pesquisa de IHC no Brasil relaciona-se com os programas de educação do país (GASPARINI et al., 2015). Em Gasparini et al. (2016), o trabalho desenvolvido buscou responder qual a influência das publicações do IHC nas publicações do próprio IHC, analisando entre outros pontos, como os números de citações de/para artigos do IHC se distribuía ao longo das edições do evento. Com estes resultados, foram possíveis diversas reflexões sobre a comunidade brasileira de IHC.

Durante um painel dentro do evento IHC 2012, a comunidade brasileira de IHC foi reunida para discutir diversos desafios da área, sendo que foram selecionados cinco grandes desafios para serem realizados durantes os próximos dez anos e publicados em forma de um relatório chamado GranDIHC-BR (SBC, 2012). Bueno et al. (2016) buscou saber como as pesquisas em IHC no Brasil estão avançando em relação a estes desafios analisando os tópicos de pesquisa apresentados em publicações da área de IHC no Brasil. Neste trabalho, o foco foi estendido, já que não se restringiu apenas ao Simpósio IHC, mas também considerou todos os outros eventos e os periódicos que os autores de IHC tiveram publicações para a geração da rede de colaboração científica.

2.2 Bibliometria, Cienciometria e Infometria

A Bibliometria começou focada na medida de livros (quantidade de edições e exemplares, quantidade de palavras contidas nos livros, espaço ocupado pelos livros nas bibliotecas), e com o tempo, foi tornando-se uma ferramenta de estudo importante para outros tipos de produções bibliográficas, como artigos de periódicos por exemplo, podendo dessa forma, estudar a produtividade dos autores e realizar o estudo das citações (ARAUJO, 2006). Conforme Alvarado (1984), a Bibliometria pode ser definida como a aplicação de métodos matemáticos e estatísticos a livros e outros meios de comunicação escrita.

Não é algo novo usar métodos estatísticos e matemáticos para mapear informações usando para isso, registros bibliográficos de diversos tipos de documentos como livros, periódicos e artigos, porém, só no século XX que este processo ganhou um maior aprofundamento e legitimidade (SANTOS; KOBASHI, 2009). Muitos confundem ainda o termo com a Bibliografia, porém, segundo Nicholas e Ritchie (1978), a principal diferença entre a Bibliometria e a tradicional Bibliografia é que a Bibliometria utiliza mais métodos quantitativos do que discursivos.

É importante definir as diferenças entre Bibliometria, Cientometria (também denominada Cienciometria), e a Infometria; Uma definição inicial é apresentada a seguir:

- **Bibliometria:** Pritchard (1969) definiu a bibliometria como um conjunto de métodos e técnicas quantitativas para gerenciar bibliotecas e instituições com o processamento das informações.
- **Cienciometria:** Conforme Price (1969), a cienciometria é o ramo da ciência da informação que procura estudar aspectos quantitativos da atividade científica, seja como uma disciplina, ou como uma atividade econômica.
- **Infometria:** De acordo com Polanco (1995):

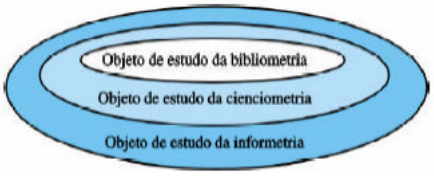
  > [...] a infometria comporta uma síntese da bibliometria e da cientometria, mas também como Brookes destacou tão bem, ela significa uma abertura ao estudo matemático da informação e sobre suas formas documentárias (Ciência social da informação) seja eletrônica ou física [...]

Já a Federação Internacional de Documentação (FID) propôs o termo Infometria

como um conjunto de métricas relativas à informação, onde aborda tanto a bibliometria quanto a cienciometria (EGGHE; ROUSSEAU, 1990).

McGrath (1989) caracterizou de forma concisa os três tipos de estudo, onde eles são subdisciplinas que se assemelham por serem métodos quantitativos, porém, são diferentes em relação aos objetos de estudo, variáveis, métodos específicos e os objetivos. Por exemplo, na bibliometria os objetos de estudo são os livros, documentos, revistas, artigos e usuários; Já na cienciometria são as disciplinas, assuntos, áreas e campos; e na infometria seriam as palavras, documentos e base de dados (MACIAS-CHAPULA, 1998). Na Figura 2.1 é apresentada a abrangência dos objetos de estudo de cada um dos métodos.

Figura 2.1 – Abrangência dos objetos de estudo da bibliometria, cienciometria e informetria.



Fonte: Pacheco and Kern (2001).

Apesar dessas classificações, os artigos geralmente categorizam por Bibliometria, usando a abordagem de que a Bibliometria usa um nicho específico que são os livros, artigos, revistas, journals, entre outros. Como neste trabalho as métricas bibliométricas serão amplamente utilizadas com base nas publicações dos autores mais prolíficos, definiremos melhor como é feito esse tipo de análise onde a Bibliometria aborda três leis básicas (SANTOS; KOBASHI, 2009; ARAUJO, 2006; ALVARADO, 1984):

1. **Lei de Bradford (Lei da Dispersão dos periódicos):** Descreve uma distribuição das publicações periódicas em uma área específica ou tema específico, dessa forma, mensurando o grau de atração de periódicos sobre determinada temática.

2. **Lei de Zipf (Lei do Mínimo Esforço):** Descreve a frequência no uso de palavras em um determinado texto, ou seja, mensura a quantidade de ocorrência do aparecimento das palavras em diversos textos, gerando uma lista ordenada de termos de uma determinada temática, utilizada para verificar qual tema científico é tratado nas publicações.

3. **Lei de Lotka (Lei do Quadrado Inverso):** Descreve a produtividade dos autores de artigos científicos por meio de um modelo de distribuição de tamanho-frequência

da produtividade dos autores em um conjunto de publicações. Assim, é importante ressaltar que essa lei se aplica a grandes volumes de publicações.

Outras definições também fazem parte do estudo das métricas bibliométricas:

- **Lei de Goffman**: Descreve a difusão da comunicação escrita como uma propagação de ideias dentro de uma comunidade. A teoria de Goffman explica que a propagação de ideias dentro de uma determinada comunidade é um fenômeno similar à transmissão de doenças infecciosas (epidêmicas) onde os autores são as pessoas, e as ideias seriam as doenças (MCGRATH, 1989).

- **Lei de Elitismo (Price):** Descreve que o número de membros da elite corresponde à raiz quadrada do número total de autores, e a metade do total da produção é considerado o critério para decidir se a elite é produtiva ou não. Segundo esta Lei, o número de autores que representaria a elite (a raiz quadrada do número total de autores) é creditada por metade de todas as contribuições (GUEDES; BORSCHIVER, 2005).

- **Obsolescência/Vida média/Idade da literatura:** Descreve a queda da utilidade de informações no decorrer do tempo. Dessa forma, é investigado o tempo em que a literatura em uma determinada área do conhecimento torna-se pouco utilizada, ou seja, o declínio do uso de determinada literatura (ARAO, 2014).

A Bibliometria é um conjunto de métodos de estudo que auxiliam e complementam a análise de redes sociais, onde se tem como principais conceitos de estudo, as entidades e relacionamentos retirados de publicações tais como os artigos científicos por exemplo (PRITCHARD, 1969). Assim, através de diversas análises feitas usando a Bibliometria, podemos retirar informações relevantes por meio destes métodos, além de identificar quais componentes serão usados para formar a rede de colaboração científica deste trabalho.

2.3 Análise de Redes Sociais (ARS)

O estudo sobre análise de redes sociais (ARS) tem crescido nos últimos anos, causado pela adoção em massa do uso de smartphones e dispositivos móveis conectados à internet

(GABARDO, 2015). A quantidade de informações a ser processada atualmente vindo das mídias sociais é muito grande e, se houver o tratamento adequado destes dados, pode se retirar informações de grande valia para as pessoas, empresas, e estudos científicos.

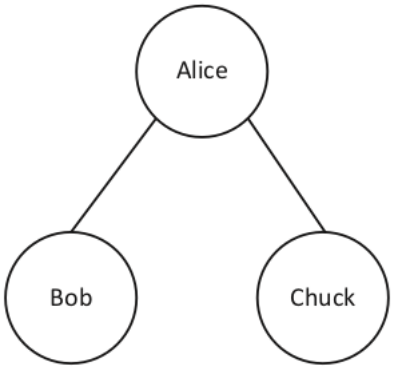
Antes de tudo, é importante deixar explícita a diferença entre mídias sociais e redes sociais, onde anteriormente as mídias sociais eram caracterizadas como uma forma de difundir uma informação de forma descentralizada, nos meios de comunicação em massa, ou seja, é o meio que determinada rede social utiliza para se comunicar, porém, hoje elas são vistas como redes de relacionamento virtuais usadas para relacionar as pessoas, e ao mesmo tempo, divulgar conteúdo tais como: Facebook, Twitter, Instagram, etc (BURKE, 2013; CIRIBELI; PAIVA, 2011).

As mídias sociais não deixam de ser redes sociais, porém, o conceito de redes sociais é muito mais abrangente, já que o termo rede social se refere a pessoas e suas conexões, ou seja, é toda rede formada por indivíduos (ou algo que possa ser individualizado) com um determinado grau de relacionamento, e isto pode ser aplicado tanto na vida real, quanto virtualmente. (GABARDO, 2015; GOLDBECK, 2015).

As mídias sociais permitem que os usuários criem contas e estabeleçam ligações entre si, gerando diversas conexões que podem ter o foco em contatos profissionais, amizades, relacionamentos amorosos, pesquisas, dentre outros (CIRIBELI; PAIVA, 2011). Já uma rede social não necessariamente online, seria por exemplo, uma equipe de desenvolvimento dentro de uma empresa de software, onde existe um grupo de pessoas que se relacionam entre si, podendo ter um relacionamento mais forte dependendo da sua função dentro da empresa, ou até mesmo ser de grau nulo. E essas relações e comportamentos não precisam necessariamente de um software para ser modelados e/ou mapeados, já que entre as tarefas está identificar apenas quem são as entidades da rede social (e.g. Pessoas) e os relacionamentos (e.g. ligações entre as pessoas).

Grande parte das análises realizadas nas redes sociais envolve o uso de suas imagens (visualização da rede social) onde cada círculo representa uma pessoa, e uma relação entre essas pessoas é representada por uma linha (GOLDBECK, 2015). A Figura 2.2 apresenta um exemplo de visualização de uma rede social, onde têm-se 3 pessoas representadas pelos 3 círculos chamadas de Alice, Bob e Chuck e as linhas representam as conexões entre Alice e Bob, e entre Alice e Chuck indicando que Alice tem conexão com ambos; porém, Chuck e Bob não possuem ligação, logo, não estão conectados entre si.

Figura 2.2 – Visualização de uma pequena rede social.



Fonte: Goldbeck (2015).

A maioria das redes sociais são muito mais extensas em escala do que este exemplo, como pode ser visto na Figura 2.3, em que a rede social do NiCHE (*Network in Canadian History & Environment*) quando começou e tinha em torno de 340 seguidores no Twitter. Cada ícone tem o tamanho de acordo com o número de seguidores que cada usuário tem. As arestas do grafo representam as conexões entre os usuários do Twitter que seguem um ao outro.

Figura 2.3 – Visualização de uma grande rede social.

Fonte: Turkel (2011).

Uma Rede Social é um estrutura social composta por indivíduos ou organizações dos quais os vértices, também chamados de "nós" são conectados por um ou

mais tipos de relacionamento de interdependência, por exemplo, amizade, trabalho, crença (DIGIAMPETRI; SILVA, 2011).

Já os grafos são estruturas de dados aplicadas em diversos problemas de várias áreas de estudo. De acordo com Cormen, Leserson e Rivest (2001), um grafo consiste em um conjunto V de vértices (ou nós) e um conjunto E de arestas. Cada aresta conecta dois vértices, e este grafo pode ser direcionado ou não, onde no grafo direcionado (também chamado de dígrafo) existe uma direção nas arestas. Em um grafo não direcionado não há ordem de relação entre os nós conectados pela aresta.

Dessa forma, a estrutura de dados por meio de grafos é uma das mais adequadas para realizar a representação computacional das redes sociais (DIGIAMPETRI; SILVA, 2011). A rede social é modelada de forma que os nós podem representar os atores e as arestas a relação entre estes autores, formando diversas características, entre elas, a centralidade, que refere-se à posição de um nó na estrutura de um grafo mostrando assim a importância e a relação do nó com os outros indivíduos (CHATTI et al., 2012).

2.3.1 Redes de Colaboração Científica

Quando existe um conjunto de pessoas ou grupos que possuem conexões de algum tipo com um ou mais integrantes de uma rede, pode ser considerado uma rede de colaboração onde o grupo busca trabalhar de forma cooperativa procurando o desenvolvimento de esforços e desenvolvimento coordenado a fim de atingir as metas em comum (NEWMAN, 2004; WEISZ; ROCO, 1996). Entre as redes de colaboração existentes, as que são formadas no âmbito científico são as Redes de Colaboração Científica formada pelos acadêmicos, professores e pesquisadores.

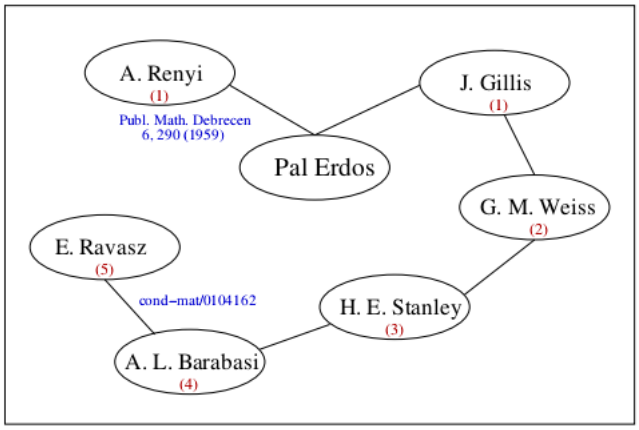
Tal colaboração pode ser formada desde uma orientação ou discussão de ideias até mesmo uma participação ativa em um projeto de pesquisa específico. O pesquisador pode ser considerado um colaborador ou até mesmo aparecer como coautor por ter fornecido um material ou auxiliado no trabalho de alguma forma (BALANCIERI, 2004). A coautoria na pesquisa científica auxilia no intercâmbio, nas relações entre autores em uma determinada área do conhecimento e nas experiências interdisciplinares (SILVA; BARBOSA; DUARTE, 2012).

Nestas redes, os vértices (ou nós) são os pesquisadores e as arestas, a cola-

boração científica, em que o relacionamento entre dois pesquisadores é a coautoria em uma ou mais publicações ou até mesmo, uma orientação de um projeto de pesquisa, ou seja, uma rede de colaboração direta, e não indireta como por exemplo, a rede de citações (NEWMAN, 2001).

No trabalho de Barabási et al. (2001), foi mostrado o conceito do número de Erdős no qual teve o nome inspirado em um dos matemáticos mais prolíficos de todos os tempos Paul Erdős, o qual escreveu mais de 1400 artigos com mais de 500 coautores. O número de Erdős define o valor zero para o autor Erdős, valor um para os coautores de uma produção com Erdős, o valor dois para os coautores desses coautores, e assim por diante. Autores com nenhuma ligação a Erdős ou a seus coautores recebem o número de Erdős infinito. Tendo isto como base, é formada uma rede de coautoria valorada como pode ser visto na Figura 2.4, apresentando assim, a distância entre um certo pesquisador e seus coautores. Este conceito pode ser aplicado à algum autor influente da comunidade de IHC e analisar a colaboração de cada pesquisador.

Figura 2.4 – Exemplo da rede formada pelo Número de Erdős.



Fonte: Ravasz (2001).

Conforme Balancieri (2004), entre os fatores que levam à formação das redes de colaboração científica têm-se:

- **Colaboração de formação (orientador-orientando):** Segundo o autor, anteriormente não se considerava um orientando como um colaborador. Porém, este tipo de relação é uma das colaborações mais evidentes, já que há a necessidade de uma contribuição especializada de um orientador para alcançar os objetivos da pesquisa, além do trabalho em conjunto contribuir para adquirir conhecimento e habilidades.
- **Colaboração teórica e experimental:** Experimentalistas tendem a colaborar

mais que teóricos já que os trabalhos teóricos produzem artigos com poucos coautores comparado aos trabalhos experimentais (BALANCIERI, 2004).

- **Proximidade na colaboração:** Quanto mais próximos os autores estiverem, mais provável será a realização da colaboração. Com a internet este critério foi reduzido, mas ainda existem alguns entraves como a correspondência cultural, de idioma, de interesses, de afinidades e oportunidades de colaboração (BARABáSI et al., 2001).

- **Produtividade e colaboração:** A alta produtividade (em termos de publicação) é de fato correlata com os altos níveis de colaboração. Enquanto colaborações com pesquisadores de alta produtividade tendem a aumentar a produtividade pessoal, colaborações com pesquisadores de baixa produtividade tendem a diminuir a colaboração pessoal (KATZ; MARTIN, 1997).

- **Quantidade de colaboradores inspira maior confiança:** Conforme o estudo de Nudelman e Landers (1972), o total de crédito que é dado por uma comunidade científica para um artigo com mais de um autor é em média maior que o crédito alocado para um único autor. O número de coautores também é fortemente correlacionado com o impacto de um artigo. Pesquisas por grandes grupos tendem a ter mais influência segundo Goffman e Warren (1980). Existem evidências que os artigos de coautoria internacional têm sido citados duas vezes mais do que um artigo de um único país (NARIN; WHITLOW, 1990).

- **Interdisciplinaridade:** Diversos avanços significativos surgiram da integração ou fusão de várias disciplinas que anteriormente eram consideradas separadas (PRICE, 1969).

- **Compartilhamento de Recursos:** A necessidade de compartilhar o uso de equipamentos caros e complexos motiva a colaboração entre os pesquisadores (**??**).

- **Reconhecimento pelos pares:** Os pesquisadores buscam aumentar a sua visibilidade e reconhecimento pelos seus pares (NARIN; WHITLOW, 1990).

2.3.2 Métricas utilizadas em Análises de Redes Sociais

Uma das análises mais interessantes da área de Análise de Redes Sociais (ARS) é determinar quais são os nós mais importantes (GOLDBECK, 2015). Para isso, têm-se diversas

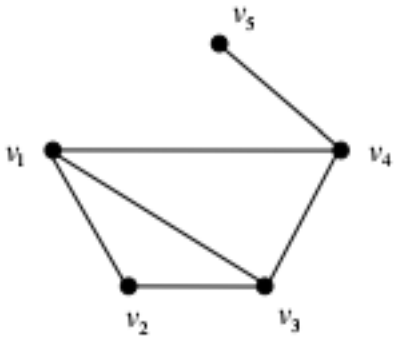
métricas para serem usadas, porém, a aplicação delas para questões específicas exigem o bom senso e o foco da análise daqueles que estiverem realizando a pesquisa com as redes sociais para poder ter uma compreensão adequada da rede (SCOTT, 2000). A seguir são apresentadas as principais métricas.

Centralidade

Centralidade é o termo usado para descrever o quão importante é um nó dentro de uma rede, podendo ser calculada de várias formas, entre elas os principais métodos são: Centralidade de Grau, Centralidade de Proximidade, Centralidade de Intermediação, e a Centralidade do Vetor Próprio (GOLDBECK, 2015).

- **Centralidade de Grau:** Contagem do número de conexões que um vértice possui. Ainda assim, podem-se usar a Centralidade Relativa de Grau, onde tem-se que pegar o coeficiente da centralidade que é o número de arestas incidentes ao vértice escolhido e dividir pelo número de nós menos 1 para se ter uma medida mais concisa. Na Figura 2.5 é mostrado um exemplo de um pequeno grafo onde os vértices v1, v3 e v4 são os mais centrais seguindo esta medida (FREITAS, 2010).

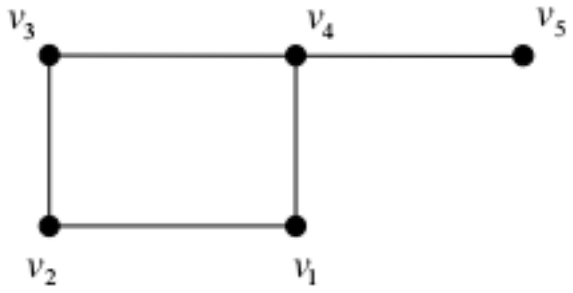
Figura 2.5 – Os vértices v1, v3 e v4 são os mais centrais segundo a centralidade de grau.



Fonte: Freitas (2010).

- **Centralidade de Proximidade:** A centralidade de proximidade mede a proximidade de um nó em relação ao grafo inteiro baseado na soma das distâncias do vértice escolhido aos demais vértices do grafo. Esta medida só é possível ser calculada para grafos conexos. Na Figura 2.6 é apresentado um exemplo de centralidade por proximidade (JúNIOR, 2016).

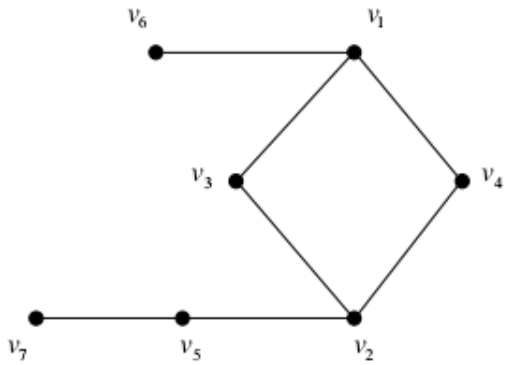
Figura 2.6 – Grafo G onde v4 é o vértice mais central segundo a centralidade de proximidade.



Fonte: Freitas (2010).

- **Centralidade de Intermediação:** Este termo foi introduzido por Freeman (1977) propondo o conceito de intermediação parcial de um vértice em uma rede para então assim, chegar a um valor que pudesse medir a centralidade deste vértice. Este valor quantifica o número de vezes que um nó age como nó intermediário ao longo do caminho mais curto entre outros dois nós (GOLDBECK, 2015). Dessa forma, ela quantifica o controle que um indivíduo tem sobre os outros indivíduos dentro de uma rede social, como podem-se ver na Figura 2.7.

Figura 2.7 – Grafo G onde v2 é o vértice mais central segundo a centralidade de intermediação.

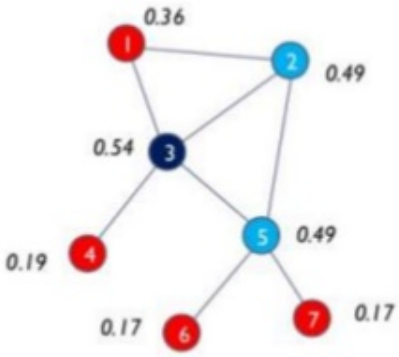


Fonte: Freitas (2010).

- **Centralidade do Vetor Próprio:** Medida de influência de um nó em uma rede. Atribui pontuações relativas aos diversos nós da rede, ou seja, a centralidade do vetor próprio é proporcional à soma das centralidades dos seus vizinhos, seguindo o conceito de que as ligações para os nós com alto valor contribuem mais para a pontuação do nó em si do que as ligações iguais a nós com baixa pontuação (JúNIOR, 2016). A Figura 2.8 apresenta o grafo com a centralidade do vetor próprio.

Centralidade do vetor próprio é uma versão mais sofisticada do grau de centralidade onde a centralidade de um nó não depende só do número de incidentes de ligações sobre o nó, mas também a qualidade dessas relações. Este fator de qualidade é determinado pelos vetores próprios da matriz de adjacência da rede.

Figura 2.8 – Grafo G onde o vértice 3 é o mais central segundo a centralidade do vetor próprio.



Fonte: Júnior (2016).

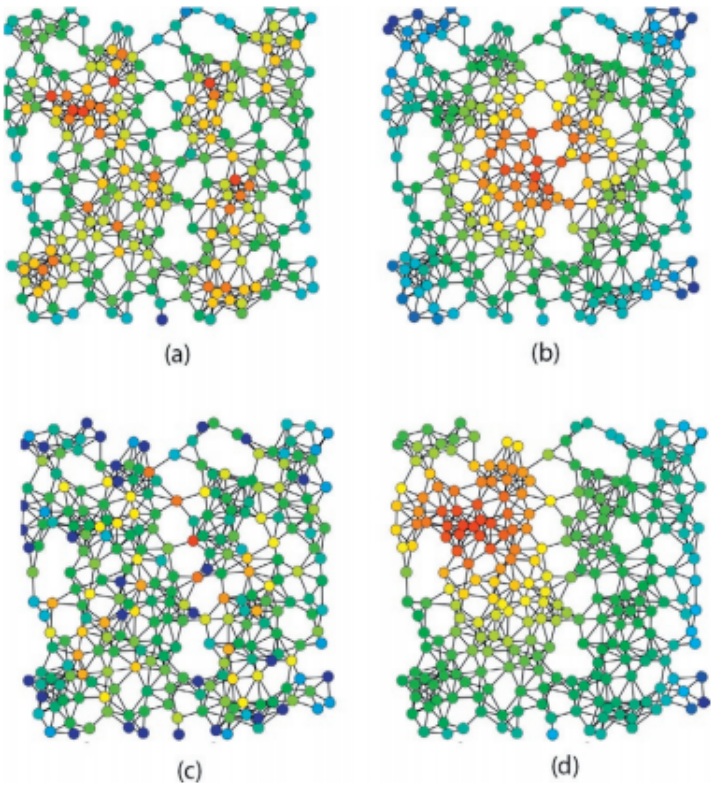
Na Figura 2.9 é apresentado um exemplo com todas as 4 centralidades calculadas para o mesmo grafo. A cor indica a centralidade de acordo com diferentes medidas onde os nós vermelhos são os mais centrais, e o nós azuis os menos centrais.

PageRank

O *PageRank* é uma avaliação da relevância de uma determinada pessoa em uma rede; Para calcular esta medida, considera-se basicamente a quantidade, qualidade e contexto de links que uma pessoa recebe e faz (FáBIORICOTTA, 2016). É uma medida muito utilizada no ranqueamento de páginas web como pode-se ver no exemplo da Figura 2.10, porém, o método também pode ser aplicado a redes sociais. Por exemplo: você joga futebol e quer saber o quão bem você está jogando. Se vários colegas seus (jogadores medianos) falarem que você joga bem não importa muito, pois eles não têm um credenciamento para falar de futebol. Por outro lado, se o Pelé e a Marta falarem que você joga bem, a relevância da sua habilidade de jogar futebol aumenta consideravelmente.

O nó C tem um valor de PageRank mais elevado do que o nó E, apesar de existirem poucas ligações para C. Mas ligação para C vem de um nó importante e, portanto, tem um valor elevado.

Figura 2.9 – A mesma rede é mostrada 4 vezes: (a) Centralidade de Grau, (b) Centralidade de Proximidade, (c) Centralidade de Intermediação, (d) Centralidade de Vetor Próprio.



Fonte: Júnior (2016).

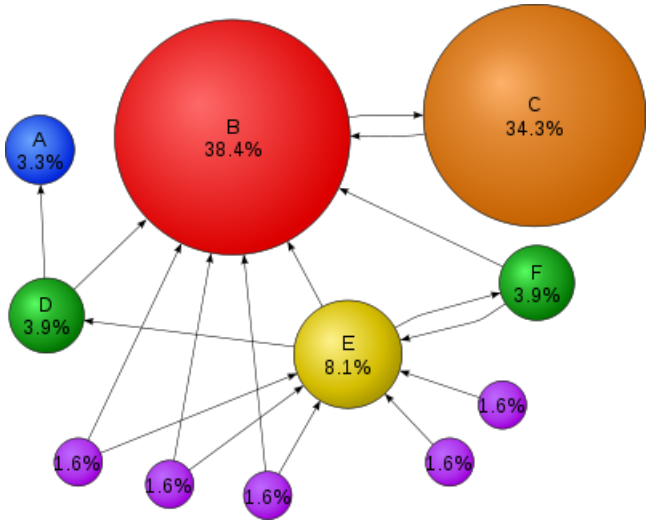
2.4 Extração de Conhecimento (KDD)

Os avanços na internet e meio de comunicação entre as pessoas estão produzindo conjuntos massivos de informação no comércio e uma diversidade de disciplinas científicas e tende a crescer cada vez mais conforme a tecnologia e a computação avançam (TAN; STEINBACH; KUMAR, 2009). O KDD (*Knowledge Discovery in Database*) ou descoberta/extração de conhecimento é um processo usado para identificar padrões válidos em analise de grandes conjuntos de dados, podendo encontrar informações importantes que podem auxiliar na tomada de decisão de diversos ramos como a lucratividade de uma empresa, por exemplo (BUENO; VIANA, 2012).

Qualquer banco de dados pode ser usado, desde que antes seja realizado uma limpeza nos dados de forma que fiquem somente os dados mais importantes e necessários. Todo o processo KDD é composto por cinco fases: seleção, pré-processamento, transformação e mineração dos dados e só então, a análise dos resultados.

A seleção busca decidir quais os conjuntos de dados serão relevantes para que sejam gerados resultados com informações úteis, e então, com os dados selecionados, entra

Figura 2.10 – Métrica PageRank para os nós de uma rede simples, expressa em percentagens.



Fonte: Domínio público.

a fase de pré-processamento, em que acontece a limpeza dos dados e a seleção de atributos (BUENO; VIANA, 2012). Na fase de transformação dos dados, os dados considerados importantes que passaram pelo pré-processamento são modificados de forma que a etapa de Mineração dos dados seja realizada, que será onde os dados finalmente serão lidos e interpretados fazendo com que os dados sejam transformados em informações relevantes (CARVALHO, 2005). E por fim, a última fase é feita a análise dos resultados, onde poderão surgir padrões, relacionamentos e descobertas de novos fatos podendo ser usado para novas pesquisas e otimizações, por exemplo.

Não existe um consenso entre a definição de KDD e *Data Mining* (Mineração de Dados) já que muitos autores como Rezende (2005), Wang (2005) e Han et al. (2006) consideram KDD e *Data Mining* sinônimos, enquanto outros tais como Cios et al. (2007) e Fayyad (1996) definem *Data Mining* como um processo que faz parte do KDD. Porém todos concordam que o processo de mineração deve ser iterativo, interativo e dividido em fases. De acordo com Fayyad (1996), o KDD é um processo não trivial de identificação de novos padrões válidos, úteis e compreensíveis. Sabendo disso, na próxima seção é explicada de forma geral o conceito de *Data Mining*, e quais modelos e técnicas podem ser usados. Neste trabalho adotou-se um padrão similar ao usado no processo de extração de conhecimento usando algumas de suas fases como Seleção de Dados, Pré-processamento dos dados, Tranformação dos Dados e a Análise dos Resultados. A fase de Mineração de Dados não foi tão útil neste momento, ficando para um trabalho futuro.

2.4.1 *Data Mining* (Mineração de Dados)

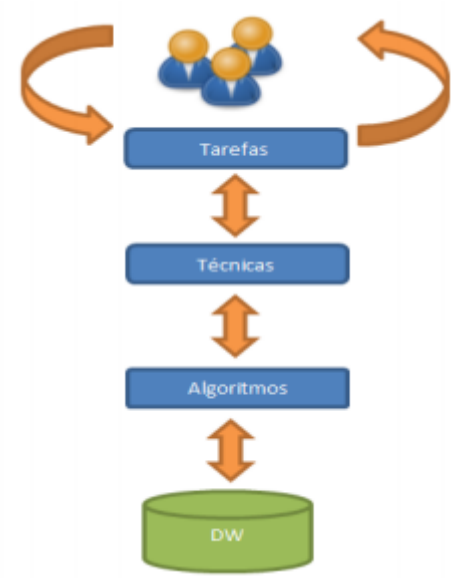
A Mineração de Dados é um conjunto de técnicas automáticas de exploração de uma grande quantidade de dados com o objetivo de descobrir novos padrões e relações que não seriam descobertas facilmente apenas com análises a olho nu (AMORIM, 2006). É na fase de Mineração de Dados do KDD que tudo acontece, pois os dados após serem transformados, serão lidos e interpretados para se retirar informações que são indicadas por meio de regras que são interpretadas via força bruta (BUENO; VIANA, 2012).

De acordo com Carvalho (2005), embora as técnicas de Mineração de Dados sejam antigas, só atualmente elas têm sido usadas para exploração de dados, por diversos motivos, tais como:

- **O volume de dados disponível é enorme -** A mineração de dados só é aplicada a uma quantidade muito grande de dados, e tais dados em massa têm sido gerados por grandes empresas de telefonia, bancos, comércio eletrônico, entre outros;
- **Os dados estão sendo organizados -** Os dados de diversas fontes têm sido organizados e padronizados possibilitando a mineração de dados que precisa ter os banco de dados limpos, padronizados e organizados;
- **Os recursos computacionais estão cada vez mais potentes -** O avanço tecnológico trouxe consigo o aumento da potência computacional, e consequentemente à queda do custo dos computadores facilitando a mineração de dados que necessita de um poder computacional alto para ser operado;
- **Banco de dados distribuídos -** A área de banco de dados sofreu diversos avanços, entre eles os bancos de dados distribuídos que auxiliam em muito à mineração de dados;
- **Competição empresarial exige técnicas modernas de decisão -** Diversas empresas atualmente estão cada vez mais competitivas buscando novas alternativas de lucrarem mais e inovarem em relação às concorrentes, e estas empresas sempre mantiveram uma grande quantidade de informação em seus bancos de dados, então naturalmente elas tiveram que usar a mineração de dados para auxiliar no apoio de tomada de decisões.

- **Programas comerciais de mineração de dados já podem ser adquiridos** - Alguns pacotes já podem ser encontrados no comércio, contendo algumas das técnicas de mineração de dados, porém, grande parte ainda se encontra em campo acadêmico.

Figura 2.11 – Interação entre os elementos do *Data Mining*



Fonte: Bueno (2012).

Há um diferença importante entre o que é uma tarefa, e o que é uma técnica de mineração de dados. Uma tarefa pode ser definida como uma especificação do que se quer buscar nos dados, quais as regularidades ou padrões são os objetivos, enquanto que, a técnica de mineração consiste na especificação dos métodos que garantem descobrir estes padrões que interessam, como podem-se ver no diagrama da Figura 2.11, as tarefas serão utilizadas pelas técnicas para que seja realizada a Mineração de Dados (BUENO; VIANA, 2012). Dentro da Mineração de Dados, três áreas são consideradas com maior expressão: Estatística, Aprendizado de Máquina e Banco de Dados como podem-se ver na Figura 2.12.

Técnicas

A Mineração de Dados é composta por diversas técnicas, sendo as mais importantes que são a Classificação, Estimativa, Previsão, Associação, Agrupamento (CARVALHO, 2005):

1. **Classificação:** Esta técnica visa identificar qual classe um determinado item pertence. O modelo analisa os dados e é capaz de dizer qual categoria um novo item se

Figura 2.12 – Principais recursos que consistem o *Data Mining*

Fonte: Bueno (2012).

encaixa. Segundo Camilo (2009), esta técnica pode ser usada, por exemplo, para:

(a) Determinar quando uma transação de cartão de crédito pode ser uma fraude;

(b) Diagnosticar onde uma determinada doença pode estar presente;

(c) Identificar quando uma pessoa pode ser uma ameaça para a segurança.

2. **Estimativa ou Regressão:** Ao contrário da classificação, a estimativa está relacionada a respostas contínuas. O objetivo desta técnica é determinar o valor mais provável diante dos dados do passado ou de dados de outros índices semelhantes sobre os quais ele tem conhecimento.

3. **Previsão ou Predição:** Similar a Classificação ou Estimativa, porém, ela visa descobrir uma informação futura de um determinado atributo, e a única maneira de saber se a previsão foi bem feita é aguardar o acontecimento e validar o quanto a previsão foi acertada ou não. Segundo Camilo (2009), são exemplos:

   (a) Predizer o valor de uma ação três meses adiante;

   (b) Predizer o vencedor do campeonato baseando-se na comparação das estatísticas dos times.

4. **Agrupamento ou Clusterização:** Nesta fase identifica-se os registros similares. O agrupamento (ou cluster) acontece reunindo em uma coleção os registros similares entre si, porém diferente de outros registros nos demais agrupamentos. Segundo Amorim (2006), agrupar sintomas pode gerar classes que não representem nenhuma doença explicitamente, já que doenças diferentes podem possuir os mesmos sintomas. Neste caso precisaria de um especialista para avaliar.

5. **Associação ou Afinidade:** Consiste em identificar quais atributos estão relacionados. Dessa forma, esta técnica tenta reconhecer os padrões de ocorrência simultânea de determinados eventos nos dados em análise. Um exemplo, segundo Amorim (2006), seriam os produtos que são comumente comprados em conjunto pelos consumidores em um supermercado.

2.4.2 Índice de Similaridade

Na Mineração de Dados existe uma análise para classificação e agrupamento de perfis dos conjuntos de dados, usando pares de objetos para comparar a similaridade e diversidade destes conjuntos, chamada de Índice de Similaridade, na qual é usado um coeficiente para ser calculado. Os coeficientes baseiam-se na comparação entre o número de atributos em comum para um determinado par de dados e o número total destes atributos. Estes coeficientes podem ser divididos em dois grupos considerando a ausência conjunta de alguns dados: os que consideram a ausência conjunta e os que não consideram a ausência conjunta (CARLINI-GARCIA, 1998)

Figura 2.13 – Notação de Similaridade de Jaccard

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Fonte: Produção do autor.

Dentre os coeficientes que consideram a ausência conjunta, existe o Coeficiente de Similaridade de Jaccard (CARLINI-GARCIA, 1998), no qual compara objetos de dois conjuntos de dados para verificar qual a similaridade dos mesmos. O intervalo desse índice vai de 0 a 1 e quanto maior o valor, mais similares os conjuntos são. A fórmula é dada como mostrado na Figura 2.13 , ou seja, pela interseção dos dois conjuntos dividido pela união dos dois conjuntos de dados.

A similaridade entre os conjuntos é usada para classificar agrupamentos de objetos. Ela pode ser útil para verificarmos se as amostras formam conjuntos discretos (OLIVEIRA, 2016). Neste trabalho não será classificado os agrupamentos conforme a similaridade, pois, o uso será apenas para comparar a similaridade dos pares de objetos.

2.5 Considerações sobre o Capítulo 2

Neste capítulo foram apresentadas os conjuntos de métodos e técnicas conhecidos como Bibliometria, Cienciometra e Infometria, as quais são adequadas para se trabalhar com um grande volume de dados. No contexto deste trabalho, estas técnicas serão associadas pois irão usar publicações de diversos autores de diversos meios de publicação. Também foram apresentados os conceitos que estão envolvidos com tema deste trabalho, como por exemplo a Comunidade Brasileira de IHC, da qual será retirada os dados a serem trabalhados por meio do Simpósio IHC.

Foram definidos os conceitos sobre Análise de Redes Sociais mostrando o que é uma Rede Social, como ela é associada a estrutura de dados de Grafos e qual a diferença entre os termos Mídias Sociais e Redes Sociais. Além disso também foi conceituado a rede de colaboração científica assim como as métricas de redes sociais, as quais serão muito úteis para este trabalho. Por fim, foi mostrado como a Descoberta/Extração de Conhecimento (KDD) pode ser usada para transformar dados brutos em informações relevantes para descoberta de informações.

As técnicas de KDD, em especial a Mineração de Dados, são muito úteis para a descoberta de informações em grandes bases de dados, usando a análise matemática para derivar padrões e tendências existentes nestes dados. Entretanto, para este trabalho, a fase de Mineração de Dados não será necessária, pois além de necessitar de uma quantidade maior de dados, seria necessário construir um modelo de transformação de dados para a realização da mineração, e buscar análises mais profundas como predições por exemplo, nas quais não se encaixam no objetivo inicial desse trabalho, sendo assim, foram escolhidas as seguintes fases do KDD para esta pesquisa: seleção, pré-processamento, transformação dos dados, e análise dos resultados.

3 Trabalhos Relacionados

Este capítulo apresenta os trabalhos relacionados. Foram selecionados trabalhos que envolvessem análise de redes sociais (principalmente no âmbito acadêmico e científico), uso de técnicas de mineração de dados para analisar uma grande quantidade de dados, e pesquisas realizadas sobre a comunidade brasileira de IHC.

Dentre os trabalhos selecionados, temos aqueles que fizeram um estudo baseado em exploração visual de dados, técnicas de mineração de dados e análise de redes sociais, para mapear e analisar as redes de colaboração científica, encontrar *insights* e tendências do assunto pesquisado, além de avaliar os resultados encontrados. Além disso, também foram selecionados algumas ferramentas, *frameworks* ou base de dados implementadas pelos autores para facilitar a pesquisa nessa área.

3.1 Barbosa, Silveira e Gasparini (2017)

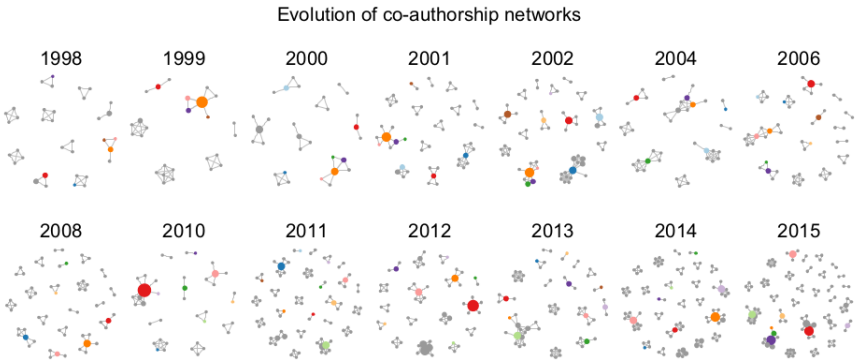
Barbosa, Silveira e Gasparini (2017) aplicaram dados e análises visuais à principal conferência de IHC dentro do Brasil, o Simpósio Brasileiro de Fatores Humanos em Sistemas Computacionais. Este trabalho explorou os dados de 340 artigos completos publicados nas 14 edições do Simpósio IHC, com o objetivo de investigar a evolução do "autoconhecimento", possibilitando o desenvolvimento de novas estratégias de pesquisa.

Entre as questões debatidas neste trabalho, estavam como o perfil dos autores se alterou ao longo do tempo, assim como a colaboração científica entre os autores do IHC evoluiu, de quais instituições e estados são as publicações do Simpósio IHC e como as referências dos trabalhos mudaram durante o tempo. Também foi visto se o idioma das publicações e das referências mudaram neste tempo e se os tópicos de pesquisa foram alterados ou evoluíram.

Como resultados, foram obtidos a evolução das co-autorias, instituições e estados mais influentes, tópicos, pesquisas e perfis dos pesquisadores ao longo do tempo. A Figura 3.1 ilustra a evolução das redes de coautoria durante todas as edições do Simpósio IHC. Este trabalho serve como referência para outros estudos sobre comuni-

dades científicas independente da área de estudo.

Figura 3.1 – Evolução das redes de coautoria: as cores representam os mesmos autores mostrados na imagem anterior



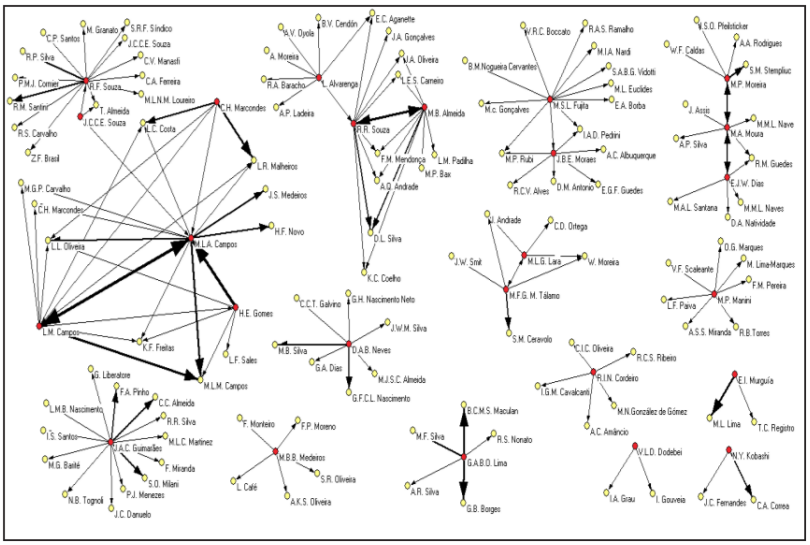
Fonte: Barbosa, Silveira e Gasparini (2017).

3.2 Silva et al. (2012)

O trabalho apresentado por da Silva et al. (2012) realizou uma análise da dinâmica das redes sociais de coautoria no campo da ciência da informação no Brasil focado no Grupo de Trabalho (GT2) "Organização e representação do conhecimento", do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB). A pesquisa caracterizou as redes quanto à produção, ao vínculo institucional, à região geográfica, aos autores mais produtivos e às redes de coautoria. Este estudo realizou uma pesquisa do tipo exploratória com base nos Anais do ENANCIB e dos currículos dos autores pesquisados e registrados na Plataforma Lattes. A pesquisa foi de cunho quantitativo e fez o uso de grafos para representar as redes.

Como resultado, percebeu-se que por conta do aumento da produção científica em autoria múltipla houve um crescimento no número de parcerias e da produção em redes de coautoria. A Figura 3.2 ilustra a Rede Social de Coautoria formada pelos autores mais produtivos do GT2/ENANCIB entre os anos de 1994 até 2001. Além disso, as técnicas de análise de redes sociais para estudo das redes de colaboração científica têm se mostrado de grande valia para mapear as redes, além de identificar e representar as relações, as especifidades e os conhecimentos entre os autores e instituições colaborativas.

Figura 3.2 – Rede Social de Coautoria dos Atores mais Produtivos do GT2/ENANCIB (1994/2011)



Fonte: Da Silva et. al (2012).

3.3 Digiampetri et al. (2012)

Digiampietri et al. (2012) produziu um banco de dados com técnicas de mineração tais como classificação, agrupamento e associação a partir de mais de um milhão de Currículos Lattes, analisando assim diversas características e relações dos currículos, formando as redes sociais a partir destes dados. Os currículos da Plataforma Lattes são uma vasta fonte de informação para a criação e análise de redes sociais de pesquisadores (Balancieri et al., 2005). O foco da pesquisa foi criar uma base de dados com os currículos minerados da Plataforma Lattes para servirem de base para a produção e análise de redes sociais de pesquisa científica. Além disso, o autor discute algumas observações que devem ser consideradas nos trabalhos que analisam currículos da Plataforma Lattes. Entre elas, o cuidado que deve ser considerado em relação à existência de currículos de homônimos na Plataforma Lattes, o preenchimento incompleto e/ou incorreto de informações nos currículos e a diferença entre o tempo de atualização dos currículos por parte dos autores.

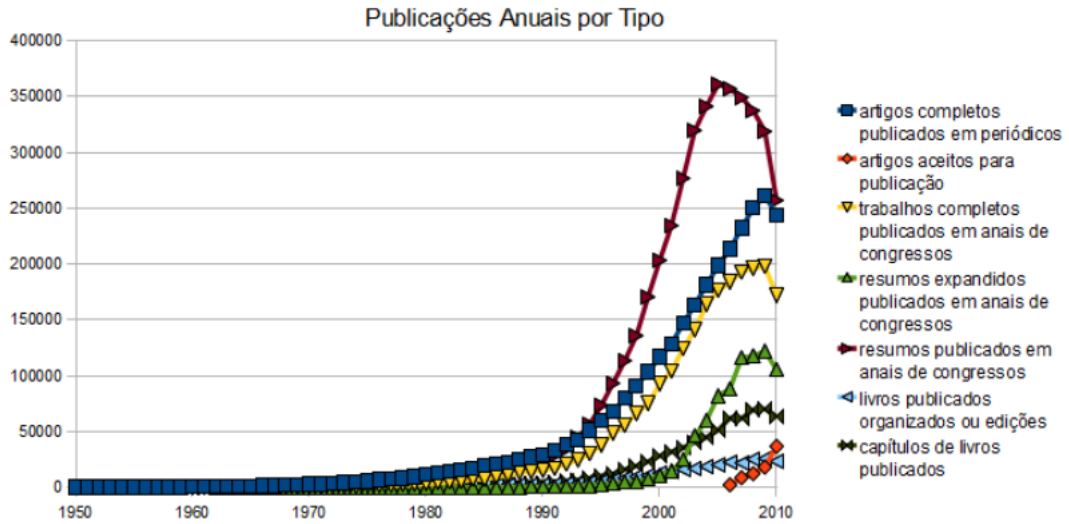
Os currículos da Plataforma Lattes utilizados para esta pesquisa foram obtidos pela internet através do comando *wget* para baixar cada um dos currículos. Cada currículo tem um identificador numérico necessário para compor a URL (*Uniform Resource Locator*) completa de cada currículo. Foram usadas duas estratégias para encontrar os identificadores dos currículos, sendo a primeira, consultas na interface de busca da Plataforma Lattes utilizando palavras-chave das (sub)áreas de conhecimento da própria plataforma. Assim,

foi implementado um *parser* para encontrar os identificadores numéricos dos currículos de cada um dos resultados da consulta. Apenas com esta primeira estratégia, foram encontrados centenas de milhares de currículos.

A partir disso, os currículos foram baixados e foi realizado a segunda estratégia que constituiu em buscar novos identificadores dentro dos currículos (por exemplo, de coautores, orientadores ou orientandos). Assim, usando as duas estratégias de busca, foram baixados um total de 1.236.548 currículos totalizando pouco mais de 16 GB em arquivos HTML (*HyperText Markup Language*). Vale ressaltar que os currículos foram baixados em 2011, a partir desta data foram feitos apenas processamentos dos dados capturados nos currículos.

Este trabalho apresentou como é possível realizar uma análise com base nos dados dos Currículos Lattes, podendo futuramente realizar uma análise muito ampla das relações entre esses dados e das redes formadas pelos seus pesquisadores. Foi verificado que os currículos possuem uma certa padronização imposta pelos formulários, além do que, grande parte da informação preenchida pelos usuários é feita de forma manual possibilitando a ocorrência de vários erros que precisam ser tratados durante o processamento e análise de currículos (DIGIAMPETRI et al., 2012). Na Figura 3.3 é mostrada a evolução do número de registros de publicações cadastrados nos currículos ao longo dos anos. Observa-se um crescimento ano a ano do número de publicações.

Figura 3.3 – Evolução no Número de Publicações no Tempo.



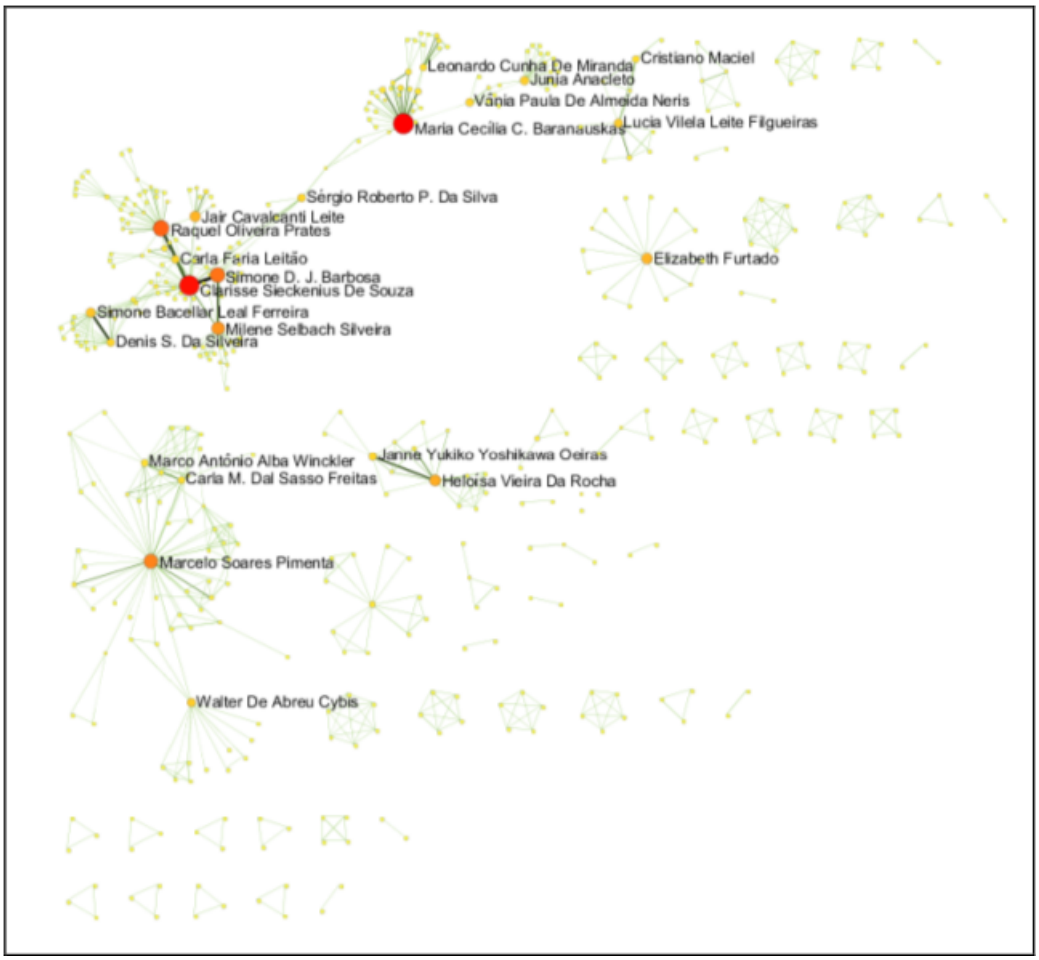
Fonte: Digiampetri et al. (2012).

3.4 Gasparini et al. (2014)

Gasparini et al. (2014) realizou a análise baseada na exploração visual orientada a dados, identificando os autores, as instituições centrais e os principais temas do Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (doravante Simpósio IHC) e as tendências, procurando mostrar a importância das redes de coautoria no estudo da produção científica. O foco da pesquisa foi analisar as redes de coautoria exclusivamente do próprio Simpósio.

Foram obtidos os Anais de todas as edições do IHC até o ano de 2013 analisando-se somente os artigos completos que foram inseridos em uma base de dados, passando por uma normalização e ajustes para melhor adequação. Foram realizadas diversas análises estatísticas sobre os autores e publicações da comunidade de IHC, e identificadas diversas redes de colaboração científica entre os pesquisadores da área.

Figura 3.4 – Rede de Coautoria dos Artigos Completos do IHC.



Fonte: Gasparini et al. (2014).

Na Figura 3.4 é apresentada a Rede de coautoria formada pelos pesquisadores

da comunidade acadêmica de IHC tomando como base os artigos completos do IHC. Os resultados mostraram a importância das redes de coautoria na produção de novos artigos em uma determinada área de pesquisa, além de permitir identificar tendências que podem estar surgindo, auxiliando os pesquisadores de IHC com novos enfoques de estudo além de permitir ter alguns *insights* sobre a comunidade de IHC (GASPARINI et al., 2014).

3.5 Maruyama e Digiampetri (2016)

No trabalho realizado por Maruyama et al. (2016) foi feito um levantamento dos atributos ou características que podem ser utilizados na predição de relacionamentos nos diversos contextos das redes sociais. Dessa forma, utilizando atributos, métodos, algoritmos e técnicas pode-se medir a possibilidade de um relacionamento ser criado. Em uma rede social é muito útil a predição de novos relacionamentos, visto que assim podem-se descobrir relações que até então eram desconhecidas ou até potencializar relações já existentes.

Primeiramente os autores realizaram uma pesquisa exploratória para identificar as principais palavras-chave relacionadas ao assunto tendo como resultado as seguintes palavras: *Link, co-authorship, prediction, social networking e scientific collaboration networking*. Após isso foi feita uma pesquisa por meio de uma Revisão Sistemática. As bibliotecas digitais *IEEExplore Digital Library* e *ACM Digital Library* foram utilizadas, e os termos selecionados anteriormente foram utilizados para encontrar artigos sobre predição de relacionamentos em qualquer tipo de rede social, assim como predição de relacionamentos em redes de colaboração científica. Para cada artigo encontrado, foi avaliado e selecionado aqueles que seguiam os critérios de inclusão e exclusão definidos no trabalho.

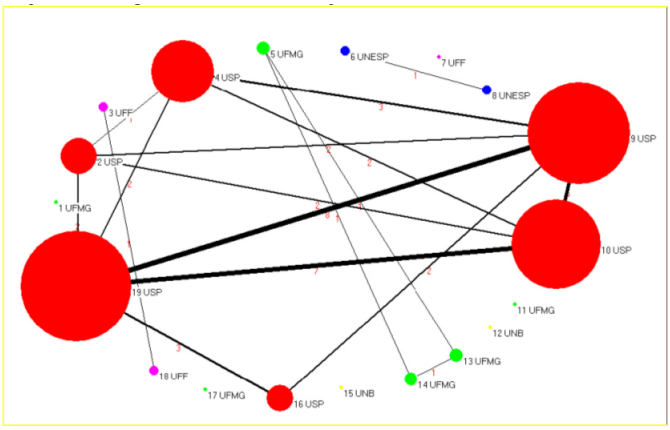
Ao final da revisão sistemática, os autores perceberam que o tema de predição de *links* para redes sociais é recente e analisando os trabalhos como um todo, existem alguns atributos considerados tradicionais neste tipo de análise como CN (*Common Neighbors*), Katz, JC (*Jaccard Coefficient*), AA (*Adamic-Adar*) e PA (*Preferential Attachment*). Também foi visto que cada rede social possui suas características únicas, sendo assim, não existe atributos ideais que satisfaçam todas as redes sociais.

3.6 Oliveira et al. (2009)

Oliveira et al. (2009) realizou uma pesquisa para proporcionar uma visão da colaboração científica entre os pesquisadores dos Programas de Pós-Graduação em Ciência da Informação do Brasil, que trabalham na área de Tratamento Temático da Informação. Ele teve como objetivo identificar, analisar e descrever a situação das redes de colaboração científica existentes, em relação à participação dos docentes. Foram levantados diversos dados observando a Plataforma Lattes como: produção de artigos, capítulos de livros e livros publicados em coautoria (intragrupo) e com pesquisadores de fora (extra grupo). Ao final, foram avaliadas as coautorias intra e extra grupo.

Nesta pesquisa foi utilizado um levantamento feito por (DANUELLO, 2007) dos artigos mais relevantes, publicados em periódicos selecionando os trabalhos com o tema Tratamento Temático da Informação. Assim, os pesquisadores encontrados foram colocados em ordem alfabética, numerados e nominados de acordo com a instituição de origem, artigos publicados em periódicos, livros e capítulos de livros. Dessa forma, avaliou-se as coautorias dentro do grupo (intra-grupo) e com pesquisadores fora do grupo (extra grupo).

Figura 3.5 – Frequência de Coautorias (Duplas e Múltiplas).



Fonte: Oliveira et. al (2009).

Com os resultados desta pesquisa, foram mapeadas as possíveis redes de colaboração científica que existem sobre o tema em questão no Brasil, a partir dos autores e instituições avaliadas. Na Figura 3.5 gerada pelo *software* Pakek, tem-se o grafo da rede formada pelas ligações entre os pesquisadores analisados, mostrando o comportamento dos pesquisadores na publicação de artigos. O ponto colorido indica o autor, e o seu tamanho indica a frequência na publicação com os pares. A espessura dos segmentos

identifica a frequência de coautorias.

3.7 Digiampetri e Silva (2011)

Digiampetri e Silva (2011) desenvolveram um *framework* para realizar análises nas redes sociais de pesquisadores combinando técnicas de Análise de Redes Sociais, Extração de Conhecimento (KDD) e Teoria dos Grafos tais como formação de redes sociais através de grafos, utilização de métricas da teoria de grafos como medidas de centralidade, seleção e classificação dos dados, agrupamento e associação, desta forma, podendo identificar, organizar, gerenciar, visualizar e sumarizar informações sobre as redes sociais de pesquisadores que tenham um currículo na Plataforma Lattes.

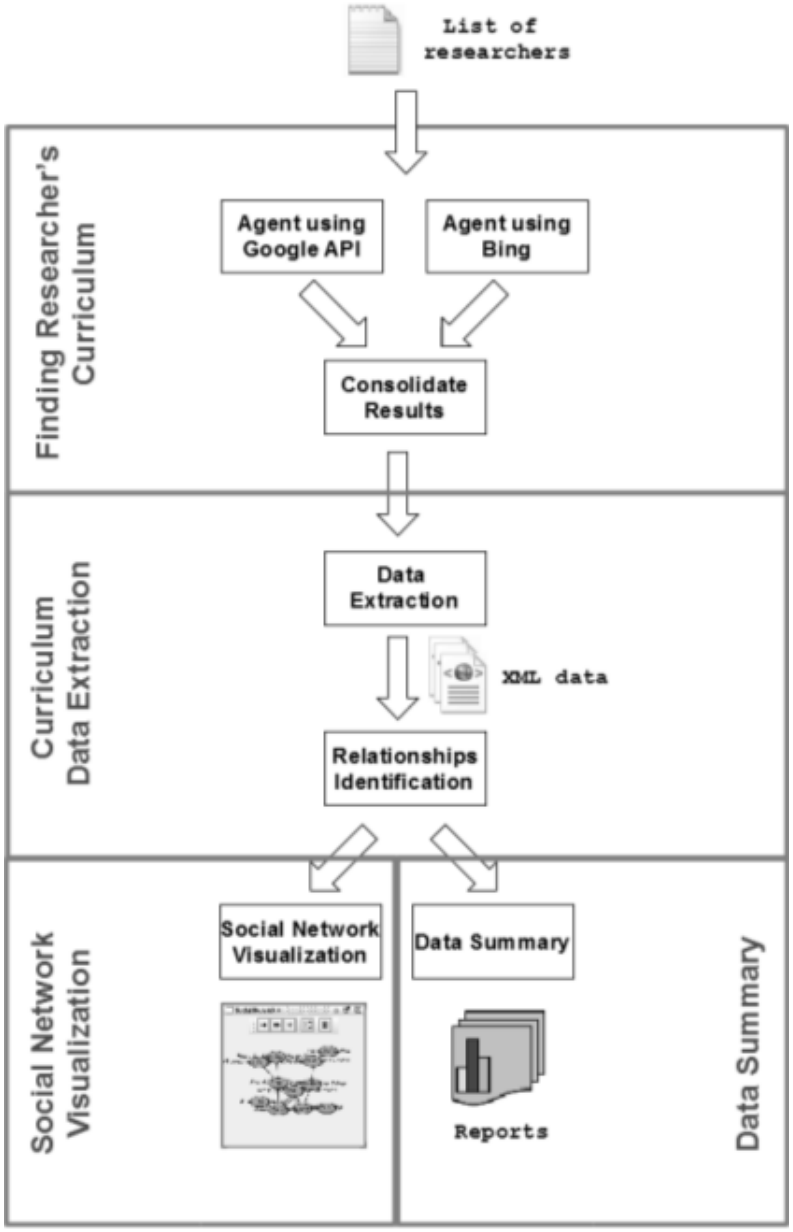
O *framework* apresentado neste trabalho foi implementado na linguagem de programação Java, e é composto por 4 sub-sistemas: (i) Encontrar Currículo dos Pesquisadores; (ii) Extração de Dados dos Currículos; (iii) Visualização das Redes Sociais; e (iv) Geração de Relatórios Resumidos, relações são demonstradas na Figura 3.6. Primeiro, é selecionado a lista de pesquisadores que serão extraídos os currículos Lattes, e assim, é usado os mecanismos de busca da Google e do Bing para buscar os currículos. Assim é extraído os currículos da Plataforma Lattes e colocado em formato XML identificando os relacionamentos dos pesquisadores. E por fim, é analisado os dados e criado a visualização da rede social formada.

O *framework* apresenta métodos de busca para encontrar os currículos Lattes de um dado pesquisador, a fim de automatizar esta tarefa que normalmente é feita manualmente. Além disso, o *framework* fornece um ambiente gráfico para permitir ao usuário visualizar a rede social e obter várias métricas para cada pesquisador, assim como para toda a rede. O sistema produz ao final vários relatórios com regras de avaliação estabelecidas pelo Comitê de Ciência da Computação da CAPES para avaliar grupos de pesquisadores de programas de pós-graduação.

3.8 ScriptLattes (2009)

Mena-Chalco, Junior e Marcondes (2009) desenvolveram uma ferramenta *open-source* para gerar relatórios acadêmicos baseados nos currículos da base de dados do Lattes. O

Figura 3.6 – Framework Overview.



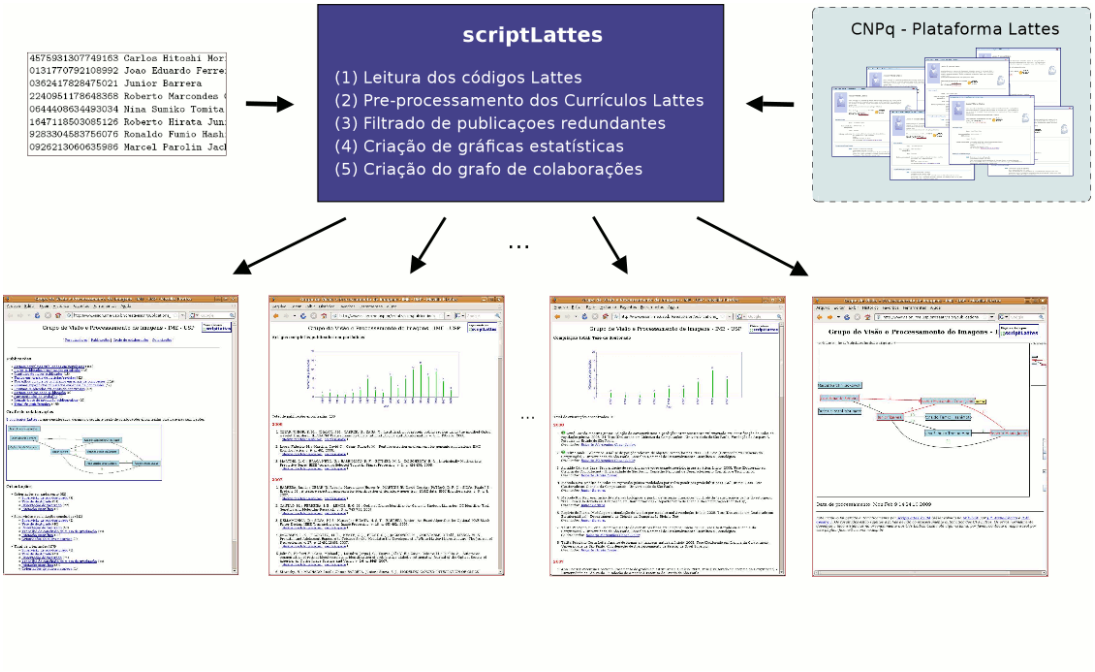
Fonte: (Digiampetri e Silva) (2011).

scriptLattes é um *script* GNU-GPL desenvolvido para a extração e compilação automática de informações de um conjunto de pesquisadores cadastrados na plataforma Lattes tais como: (1) Produções bibliográficas; (2) Produções técnicas; (3) Produções artísticas; (4) Orientações; (5) Projetos de pesquisa; (6) Prêmios e títulos; (7) Grafo de colaborações; (8) Mapa de geolocalização; (9) Coautoria e internacionalização.

A ferramenta chamada *scriptLattes* recebe os currículos Lattes em formato *HTML* de um grupo de interesse, compila as listas de produções, tratando inclusive duplicatas e similares. Assim, são geradas páginas *HTML* com listas de produções e orientações separadas por tipo e colocadas em ordem cronológica invertida. Também são

criados vários grafos de co-autoria entre os membros do grupo de interesse e um mapa de geolocalização dos membros e alunos com orientações concluídas. Porém, por conta da validação *captcha* nos Currículos Lattes, o programa não funciona de forma automática.

Figura 3.7 – Funcionamento do *scriptLattes*.



Fonte: http://scriptlattes.sourceforge.net/ (2016).

Além do benefício da própria ferramenta, a ferramenta produz a representação gráfica das redes de coautoria, e os relatórios gerados permitem analisar a produção dos grupos de pesquisa em relação à sua produção bibliográfica, técnica ou artística, orientações, participação em bancas examinadoras, eventos ou comissões julgadoras. Na Figura 3.7 é apresentado o diagrama do funcionamento da ferramenta *scriptLattes*.

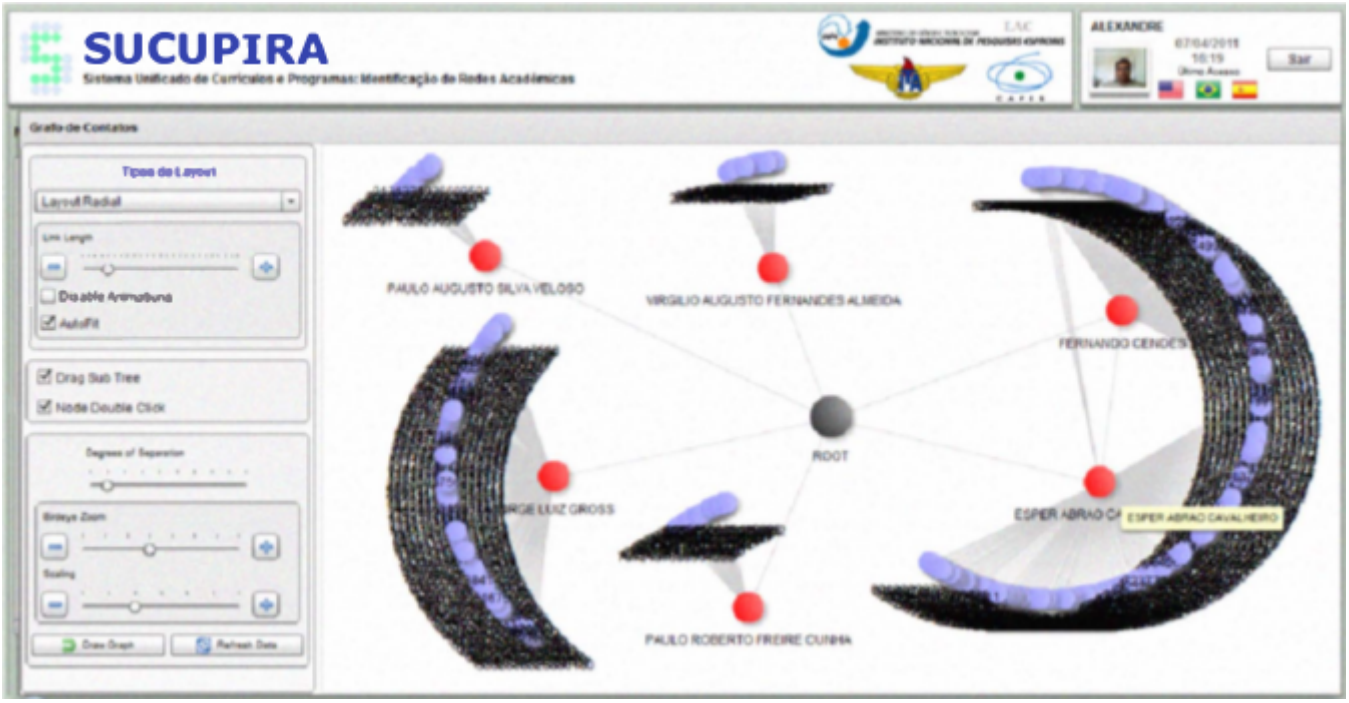
Desde a adoção dos *captchas* nos currículos da Plataforma Lattes, o *scriptLattes* não funciona de forma automática. Sendo assim, é necessário o usuário baixar cada um dos currículos Lattes, configurar o arquivo de configuração e executar o *scriptLattes*. Outra questão é que as análises são limitadas ás que a ferramenta oferece, dessa forma, ela até poderia ajudar realizar algumas análises, mas não daria total liberdade caso fosse necessário realizar alguma análise em específico, fora que também teria o trabalho manual de qualquer forma.

3.9 Sucupira (2011)

O projeto denominado SUCUPIRA (Sistema Unificado de Currículos e Programas: Identificação de Redes Acadêmicas) realizado por (ALVES; YANASSE; SOMA, 2011), financiado pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) lembra o sobrenome do falecido Professor Emérito da Universidade Federal do Rio de Janeiro, Newton Lins Buarque Sucupira que foi relator do importante Parecer 977/65 sobre a Pós-Graduação no Brasil. A principal ferramenta deste projeto é o sistema SUCUPIRA que visa ser uma ferramenta computacional automatizada e de domínio público que irá auxiliar na captura de indicadores de desempenho de docentes, pesquisadores e programas de pós-graduação. A ferramenta será usada para extração de informações da Plataforma Lattes para identificar as redes sociais acadêmicas.

O sistema SUCUPIRA foi implementado em 2011 usando a tecnologia *Adobe Flex* que suporta desenvolvimento de aplicações ricas para Internet e compatíveis com várias plataformas possuindo uma interface web. Porém, hoje conta uma nova plataforma de coleta de dados, gestão e avaliação da CAPES (Comissão de Aperfeiçoamento de Pessoal do Nível Superior), substituindo a Coleta CAPES e unificando todas as bases de dados da Pós-Graduação, incluindo Avaliação, Cadastro de Discentes e outros.

Figura 3.8 – Funcionamento do Sistema Sucupira: Grafo de Contatos de Grau 2 de Separação.



Fonte: Alves et al. (2011).

Entre as principais funcionalidades do sistema SUCUPIRA está o gerenciamento de pesquisadores definida por cada usuário do sistema, onde ele pode adicionar uma lista dos pesquisadores que ele deseja analisar e comparar. Assim, após a escolha dos pesquisadores, o sistema mostra a distribuição geográfica dos pesquisadores, o gráfico de publicações desses pesquisadores, e a visualização das redes sociais acadêmicas identificadas entre os pesquisadores. Na Figura 3.8 é apresentado o exemplo do grafo de contatos de grau 2 de separação formado pelo sistema para um dado grupo de pesquisadores.

Por enquanto, está disponível o módulo Coleta de Dados na Plataforma Sucupira e algumas ferramentas de gestão do Sistema Nacional de Pós-Graduação (SNPG), como por exemplo, solicitações de mudança de nome do curso/PPG (Programa de Pós-Graduação), mudança de área, registro de início de funcionamento, desativação de PPG. As demais ferramentas e aplicativos utilizados no acompanhamento e avaliação da pós-graduação serão gradativamente incorporados na Plataforma.

3.10 Considerações sobre o Capítulo 3

Apoiado pelos trabalhos relacionados que realizaram as gerações de redes de colaboração científica específicas para seus respectivos trabalhos e análises das mesmas, principalmente no trabalho realizado por (GASPARINI et al., 2014), que analisou as redes de coautoria do próprio Simpósio, este trabalho irá extrair os autores mais prolíficos da comunidade brasileira de IHC por meio da base de dados do Simpósio (dados já obtidos por (GASPARINI et al., 2014)). Mas no contexto deste trabalho, a pesquisa será estendida, visto que irá analisar toda a rede de colaboração científica formada por estes membros mais prolíficos da comunidade brasileira de IHC. Para tal, serão observadas todas as publicações destes autores nos diversos veículos de publicação nos últimos 10 anos.

Portanto, este trabalho visa gerar as redes de colaboração científica da comunidade brasileira de IHC, e para isso foi criado um modelo para transformação dos dados em informação (conhecimento) através das técnicas de mineração de dados passando por diversas fases (por exemplo, Pesquisa e Entendimento dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação) assim como foi feito no trabalho de (DIGIAMPETRI; SILVA, 2011). Os dados foram inicialmente capturados dos Anais do Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais selecionando os autores

que foram pesquisados na Plataforma Lattes para obtenção dos seus currículos registrados via formato *XML* (*eXtensible Markup Language).*

A partir disso, foi implementada uma ferramenta para extração automática dos dados dos currículos selecionados que foram inseridos em um banco de dados relacional para geração de tabelas com todas as informações de forma que fosse possível analisar e realizar a limpeza e padronização dos dados para que pudesse incluir essas informações em outro banco de dados não relacional, no caso, um banco de dados orientado a grafos. Após isso, foram selecionadas de análise de redes sociais para gerar as redes de colaboração científica da comunidade brasileira de IHC, analisá-las com base nestes trabalhos já feitos anteriormente.

A Tabela 3.1 compara os trabalhos relacionados de acordo como foi feito o tratamento dos dados, desde a sua obtenção até a análise dos resultados. Já na Tabela 3.2 foram analisados separadamente os trabalhos que envolveram algum tipo de desenvolvimento de ferramenta ou similar.

Tabela 3.1 – Análise Comparativa entre as Ferramentas Selecionadas.

Ferramenta	**Ano**	**Fonte de Dados**	**Técnicas usadas**	**Linguagem usada**
scriptLattes	2009	Plataforma Lattes	Teoria dos Grafos Análise Estatística dos Dados	Python
Plataforma SUCUPIRA	2011	Plataforma Lattes	Teoria dos Grafos Análise Estatística dos Dados Análise de Redes Sociais	MXML
Framework Digiampetri e Silva	2011	Plataforma Lattes	Teoria dos Grafos Análise de Redes Sociais Extração de Conhecimento	Java

As ferramentas pesquisadas foram implementadas em diversas linguagens como *Perl* (*scriptLattes* posteriomente foi atualizada para *Python*), *MXML* e *Java*. Além disso todas utilizaram técnicas parecidas tais como Teoria de Grafos, Análises de Redes Sociais, Análise Estatística dos Dados, e Mineração de Dados, porém, todas extraindo informações da Plataforma Lattes. É curioso o fato de todas terem sido implementadas na mesma época.

Tabela 3.2 – Análise Comparativa dos Trabalhos Relacionados.

Trabalhos	**Ano**	**Fonte de Dados**	**Técnicas Usadas**	**Representação dos Dados**
Silva et al.	2012	Anais do ENANCIB Plataforma Lattes	Análise Exploratória dos Dados Análise de Redes Sociais	Grafos Tabelas Gráficos
Oliveira et al.	2009	Plataforma Lattes	Análise Exploratória dos Dados Análise de Redes Sociais	Tabelas Grafos
Gasparini et al.	2014	Anais do IHC	Análise Exploratória dos Dados Análise de Redes Sociais	Grafos Tabelas Gráficos Mapa de Calor Nuvem de Tags
Digiampietri et al.	2012	Plataforma Lattes	Análise Exploratória dos Dados Análise de Redes Sociais Mineração de Dados	Tabelas Gráficos
Maruyama et al.	2016	IEEE Digital Library. ACM Digital Library ACM Digital Library	Análise Exploratória dos Dados Análise de Redes Sociais	Tabelas Gráficos
ScripLattes	2009	Plataforma Lattes	Análise Estatística dos Dados Análise de Redes Sociais	Tabelas Gráficos Grafos
Plataforma SUCUPIRA	2011	Plataforma Lattes	Análise Estatística dos Dados Análise de Redes Sociais	Tabelas Gráficos Grafos
Framework	2011	Plataforma Lattes	Teoria dos Grafos Análise de Redes Sociais Extração de Conhecimento (KDD)	Tabelas Grafos

Percebe-se pela Tabela 3.2 que a Plataforma Lattes é um acervo com uma grande quantidade de informações relevantes a serem descobertas já que é o grande alvo da maioria dos pesquisadores. Também é mostrado que boa parte dos trabalhos começaram de 2009 em diante, mostrando que com o avanço no estudo em análise de redes sociais,

recentemente causado pela internet e mídias sociais, possibilitou as pesquisas em diversas outras áreas como no âmbito científico por exemplo. A representação dos dados é a forma com que os autores usaram para representar os resultados da pesquisa e das análises.

As ferramentas que já foram implementadas mostraram o que existe até agora em relação a extração de dados da Plataforma Lattes auxiliando para saber quais tecnologias e ferramentas podem ser usadas para isso, além de demonstrar o que talvez possa ser melhorado ou o que ainda não foi feito.

Em relação aos trabalhos de forma geral, os trabalhos que envolveram a área de IHC especificamente como a pesquisa apresentada por Gasparini e colegas (2014, 2017) ajudam a mostrar quais análises podem ser feitas na rede de colaboração científica que será analisada. Enquanto que os trabalhos apresentados principalmente por Digiampetri explora como utilizar a Plataforma Lattes para obter informações para realizar análises sobre comunidades acadêmicas e como devem ser feitas, além das diversas redes sociais acadêmicas formadas pelos outros trabalhos apresentados e suas análises.

4 Trabalho desenvolvido

O trabalho desenvolvido teve como objetivo gerar e analisar as redes de colaboração científica da comunidade brasileira de estudantes, professores e pesquisadores da área de Interação Humano-Computador. Para tal, foram selecionados apenas os autores mais prolíficos de IHC com base nos anais da principal conferência brasileira da área, o Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (Simpósio IHC). Gasparini et al (2016) já havia realizado a análise dos autores mais prolíficos de todas as edições do IHC (1998 - 2015) utilizando como critério os autores que mais produziram em toda história do simpósio IHC (autores com três publicações ou mais).

Então com base nesta análise anteriormente realizada por Gasparini et al. (2016), foi re-selecionado os autores mais prolíficos usando como critério os autores que mais produziram em toda a história da conferência (autores com cinco publicações ou mais). O número escolhido de cinco ou mais publicações foi escolhido pois como a extração dos currículos da Plataforma Lattes seria de forma manual, não poderia ser um númermo muito grande, além disso, com este número, foram extraídos uma quantidade regular e suficiente para que houvesse uma base de dados concisa para realizar esta pesquisa, e também, pelo fato de que foi considerado imprescindível um autor prolífico ter ao menos cinco publicações no Simpósio IHC. Assim, foram extraídos e armazenados de forma manual cada um dos currículos dos autores em formato *XML* (*eXtensible Markup Language*) através da Plataforma Lattes.

A partir disso, foi feito o estudo mais aprofundado da linguagem de programação PHP (*PHP: Hypertext Preprocessor*, originalmente *Personal Home Page*) e da linguagem de consulta estruturada *SQL* usando o sistema de gerenciamento de banco de dados *MySQL*. Assim, foi implementado uma ferramenta de captura automática de informações com base nos currículos *XML*, na linguagem de programação *PHP*. As informações extraídas foram inseridas em um banco de dados *MySQL* dividido em diversas tabelas, separadas por categorias de acordo com o tipo de cada informação.

Tendo a base de dados devidamente pronta, foram utilizado as técnicas do KDD, entre elas, o entendimento dos dados verificando quais dados seriam utilizados

para formar a nova base e como seriam preparados. Após realizada a análise, foi feita a preparação dos dados, onde foi realizada uma extensa padronização e limpeza das informações para que não houvesse nenhuma inconsistência na base. Na preparação dos dados, dependendo de cada objetivo, foi realizado o agrupamento de informações, associações, ou limpeza de dados. Como resultado, tivemos diversas tabelas geradas para que pudessem ser realizadas análises bibliométricas e também as tabelas prontas para serem inseridas no novo banco de dados em grafos para gerar as redes de colaboração científica.

4.1 Tecnologias envolvidas e utilizadas

Esta seção apresenta as principais tecnologias estudadas para o desenvolvimento da ferramenta, e para a realização das análises das redes de colaboração científica.

4.1.1 MySQL

O *MySQL* é um sistema de gerenciamento de banco de dados (SGBD) de código aberto, o qual utiliza a linguagem *SQL* (*Structured Query Language*) como interface, sendo uma opção extremamente popular como banco de dados integrado, distribuído por milhares de *ISVs* (*Independent Software Vendor*) e *OEMs* (*Original Equipment Manufacturer*) (MYSQL, 2016).

Atualmente, é o banco de dados mais conhecido do mundo, com comprovado desempenho e facilidade de uso, tornando-se a principal opção de SGBD para aplicativos baseados na Web, usado por grandes empresas como *Facebook, Twitter, Youtube e Google* (ORACLE, 2016). O *MySQL* tem como vantagem a facil integração com o *PHP*, fazendo com que várias grandes empresas da Web utilizem o *MySQL* em aplicações de dados críticas como *Yahoo!Finance, Motorola, NASA, Silicon Graphics*, entre outras (MYSQL, 2016). Por conta de ser um dos mais confiáveis banco de dados e ser uma ferramenta *opensource*, ela foi escolhida para ser usada neste trabalho.

As principais características do *MySQL* são: Portabilidade (suporta praticamente qualquer plataforma atual); possui um desempenho excelente comparado a outros SGBD (Sistema de Gerenciamento de Banco de Dados) e estabilidade do sistema (FERREIRA; JúNIOR, 2012); suporta *Unicode, Full Text Indexes*, Replicação, *Hot Backup*,

GIS (*Geographic Information System*), *OLAP* (*Online Analytical Processing*) e outros recursos de banco de dados; compatibilidade (diversos módulos para várias linguagens de programação, como *Delphi, Java, C/C++, C#, Visual Basic, Python, Perl, PHP, ASP* (*Active Server Pages*) e *Ruby*).

Além disso o *MySQL* não é um sistema exigente em relação ao hardware; Suporta controle transacional e replicação facilmente configurável; Contém interfaces gráficas caso seja requerido; Tem acesso a vários *Storage Engines* como *MyISAM* (antigo mecanismo de armazenamento do MySQL para versões anteriores à 5.5), *InnoDB, Falcon, BDB* (*Berkeley DB*), *Archive, Federated, CSV* (*Comma-separated values*), *Solid*; Suporta *Triggers, Cursors* (*Non-Scrollable* e *Non-Updatable*), *Stored Procedures* e *Functions*; e é um Software Livre com licença *GPL* (*General Public License*) (Porém, se o programa que acessar o *MySQL* não for *GPL*, a licença comercial deve ser adquirida) (MYSQL,).

4.1.2 Neo4J

Os bancos de dados relacionais não possibilitam que os dados sejam representados de uma forma natural, assim, algumas pesquisas se tornam bem complexas ou impossíveis nos bancos de dados relacionais. Porém em uma estrutura de grafo, conseguimos os resultados de uma forma bem mais simplificada (ALMEIDA, 2011).

Dessa forma, surgem os bancos de dados não relacionais com diferentes formar de persistir os dados, como por exemplo, a estrutura de um grafo. O Neo4J por exemplo, é um banco de dados orientado a grafos, que permite que os dados sejam persistidos e percorridos na estrutura de um grafo. Ele ainda mantém algumas características que são comuns em bancos de dados relacionais, como controle de transações, suporte a pesquisas textuais, entre outras funcionalidades (NEO4J, 2017).

Na Figura 4.1 podemos perceber que as consultas realizadas no Neo4J podem ter uma performance significativamente maior que à mesma consulta em um banco de dados relacional (ROBINSON; WEBBER; EIFREM, 2015)

O banco de dados Neo4J foi escolhido para este trabalho por ser um banco de dados orientado a grafos e de fácil uso, além de ser *opensource*. Com ele poderão ser geradas e armazenadas as redes de colaboração científica, assim como tornar possível a visualização e análise das mesmas.

Figura 4.1 – Um exemplo de pesquisa em um banco de dados relacional versus a mesma pesquisa no Neo4j

Depth	RDBMS execution time(s)	Neo4j execution time(s)	Records returned
2	0.016	0.01	~2500
3	30.267	0.168	~110,000
4	1543.505	1.359	~600,000
5	Unfinished	2.132	~800,000

Fonte: Robinson, Webber e Emil Eifrem (2015)

4.1.3 PHP

O PHP (*PHP: Hypertext Preprocessor*, originalmente *Personal Home Page*) é uma linguagem de *script open source* de uso geral, adequada para o desenvolvimento web e que pode ser embutida dentro do *HTML* sendo uma linguagem interpretada, usada originalmente apenas para o desenvolvimento de aplicações presentes do lado servidor, gerando conteúdo dinâmico na Web (PHPMANUAL, 2016). O *PHP* pode ser instalado na maioria dos sistemas operacionais de forma gratuita e é um concorrente direto da tecnologia *ASP* da *Microsoft*.

O *PHP* também é um software livre, licenciado pela *PHP License*, uma licença incompatível com a *GPL* devido a algumas restrições no uso do termo PHP. Entre as aplicações que usam a linguagem estão o Facebook, Joomla, WordPress, Magento, entre outras. As suas principais características são: Velocidade e Robustez de processamento; Suporte à Orientação a Objetos; Portabilidade (independente de plataforma); Possui Tipagem Dinâmica; É uma ferramenta *OpenSource*; Para desenvolvimento de aplicações do tipo *Server-side* (Lado Servidor); Suporte a diversas extensões (PHP, 2016).

O *PHP* foi escolhido para realizar a implementação do *script*/ferramenta de extração por possuir uma extensão chamada *SimpleXML* no qual permite ler e manipular arquivos *XML*, os quais serão os tipos de arquivos que serão extraídos da Plataforma Lattes, além de ser *opensource* e de fácil uso.

4.1.4 Python

A linguagem de programação *Python* foi projetada com a filosofia de enfatizar a importância do esforço do programador sobre o esforço computacional. Ela prioriza por exemplo a legilibidade do código sobre a velocidade ou expressividade. Hoje possui desenvolvimento comunitário, aberto e gerenciado pela organização sem fins lucrativos *Python Software Foundation*. E apesar de várias partes da linguagem possuírem padrões e especificações formais, a linguagem como um todo não é formalmente especificada (FOUNDATION, 2017)

A linguagem é de alto nível, e tem como características por ser interpretada, de script, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte, além de combinar uma sintaxe concisa e clara com os poderosos recursos de sua biblioteca padrão e por módulos e frameworks desenvolvidos por terceiros. (VENNERS, 2003)

A linguagem Python foi utilizada para realizar algumas análises bibliométricas tal como a Similaridade de Jaccard, e também para realizar algumas transformações/padronizações de tabelas de dados tanto para a realização de análises, quanto para adequar as tabelas para serem inseridas no banco de dados Neo4J.

4.1.5 XML

O *XML* (*eXtensible Markup Language*) é um subtipo da *SGML* (*Standard Generalized Markup Language*) utilizado para descrever diversos tipos de dados, recomendado pela *W3C* (*World Wide Web Consortium*) como principal propósito a facilidade de compartilhamento de informações através da internet (W3C, 2016b).

O *XML* é um padrão com um grande número de aceitação, tendo sua origem em uma instituição de padronização das mais abertas e dinâmicas, a *W3C*. Entre as vantagens tem-se que ele é baseado em textos simples, suporta *Unicode* e pode representar estruturas de dados tais como: Listas, Árvores e Registros. Já como desvantagens apresentam-se a velocidade (grandes quantidades de informação repetidas prejudicam a velocidade); a editabilidade do *XML* é pouco intuitiva principalmente para grandes arquivos em editores *txt* por pessoas leigas. Além disso, o XML simples pode ser substituído por outros formatos mais simples como *YAML* (*Yet Another Multicolumn Layout*), *JSON* (*JavaScript Object Notation*) e *Simple Outline XML*.

As principais características são: separação do conteúdo da formatação; simplicidade e legibilidade para os humanos e para os computadores; criação de arquivos para validar estruturas (*XML Schema*); criação de um conjunto de declarações de marcação que definem os blocos de construção lícitos de um documento *XML* (*DTDs - Document Type Definition*); concentração na estrutura da informação, e não na aparência; interligação de bancos de dados distintos; possibilidade ilimitada de criação de Tags (W3C, 2016a).

O *XML* foi a extensão utilizada nos arquivos que continham os currículos dos autores, nos quais já eram baixados nesta extensão diretamente da Plataforma Lattes.

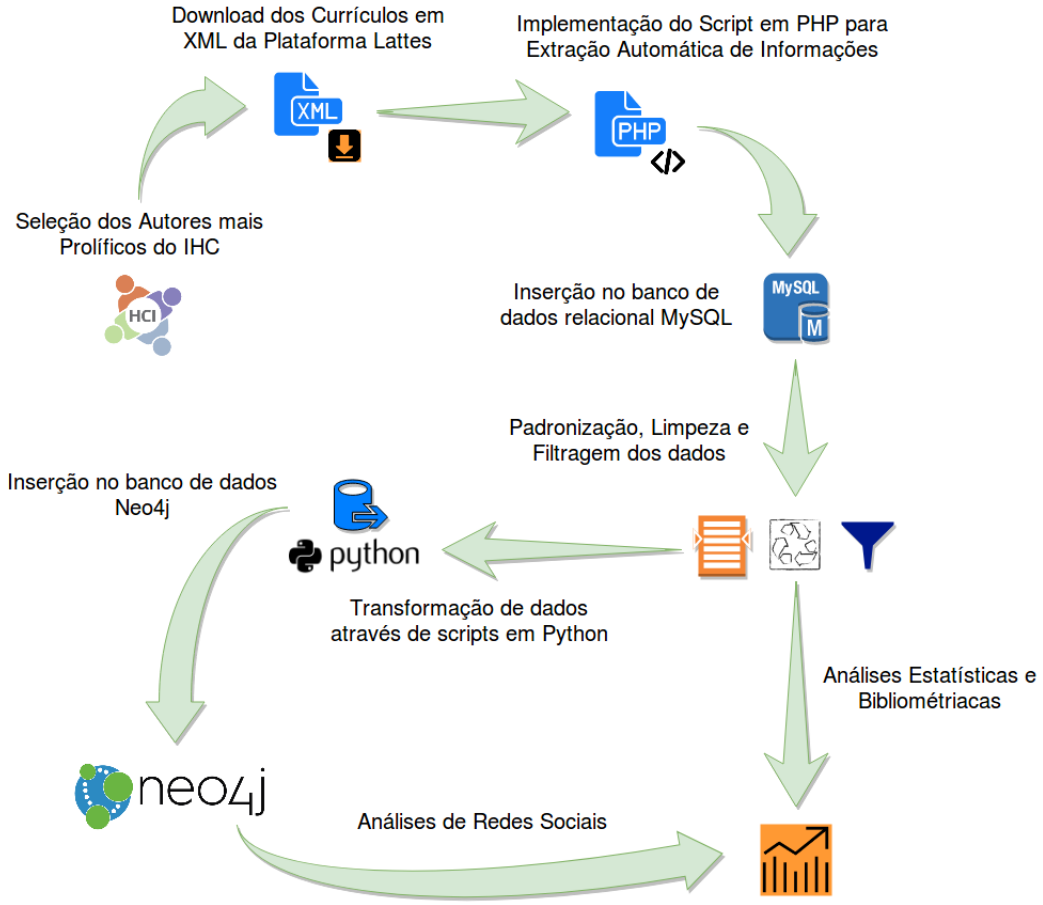
4.2 Modelo de Extração de Conhecimento (KDD)

Para a realização deste trabalho, foi realizado um modelo de extração de conhecimento com base nas técnicas mencionadas no KDD como pode ser visto na Figura 4.2, entre elas as fases de: Seleção dos Dados (Seleção e Captura dos currículos dos autores mais prolíficos de IHC), o Pré-Processamento ou Entendimento e Preparação dos Dados (Implementação do Script para extrair informações dos currículos e inserir no banco de dados para que fossem feitas a Filtragem, Limpeza e Padronização dos Dados), a Transformação dos Dados (Transformação das tabelas via *script*) em que as tabelas foram adaptadas tanto para que as análises fossem feitas, quanto para a inserção no banco de dados orientado a grafos para que pudessem ser visualizadas e analisadas as redes, e a própria Análise dos Resultados em si.

4.2.1 Seleção e Extração dos Currículos Lattes dos Autores mais Prolíficos de IHC

A seleção dos autores foi realizada usando como base os Anais do Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (Simpósio IHC), usando como critério os autores mais prolíficos dentro do evento, ou seja, que tiveram mais de cinco publicações em todas as edições do IHC entre 1998 até 2015 (até o começo desta pesquisa, 2015 era o ano mais atual dos anais). Para isso foi utilizada uma base de dados já realizada anteriormente

Figura 4.2 – Modelo de Extração de Conhecimento (KDD)



Fonte: Produção do Autor (2017)

por Gasparini et al. (2016) na qual foram selecionados os autores com mais produções em todas edições do Simpósio IHC. Desta forma, para este trabalho foram selecionados 29 autores, e o número respectivo de publicações que cada um contribuiu para o simpósio IHC.

Após ter o conhecimento de quais os autores mais prolíficos de IHC, foi realizada a extração manual de cada um dos currículos Lattes dos autores, no formato *XML* pela Plataforma Lattes. Para tal, os currículos foram acessados manualmente e realizado os *donwloads* individualmente, no período de setembro de 2016. Cada currículo lattes em *XML* é composto por várias *tags* estruturando o currículo em diversas categorias, entre as principais:

- **Dados Gerais:** Informações tais como dados pessoais, resumo do currículo *vitae*, endereço, formação acadêmica, atuações profissionais, áreas de atuação e idiomas;

- **Produções Bibliográficas:** Informações de trabalhos em eventos, artigos publicados, livros e capítulos de livros publicados, textos publicados em jornais ou revistas e demais produções bibliográficas;

- **Produções Técnicas:** Informações sobre software desenvolvidos, trabalhos técnicos produzidos e demais tipos de produções técnicas.

- **Outras Produções:** Informações das orientações concluídas e demais trabalhos realizados;

- **Dados Complementares:** Informações das participações em banca, participações em eventos e congressos, orientações em andamento e outras informações adicionais sobre instituições e cursos.

4.2.2 Implementação da Ferramenta

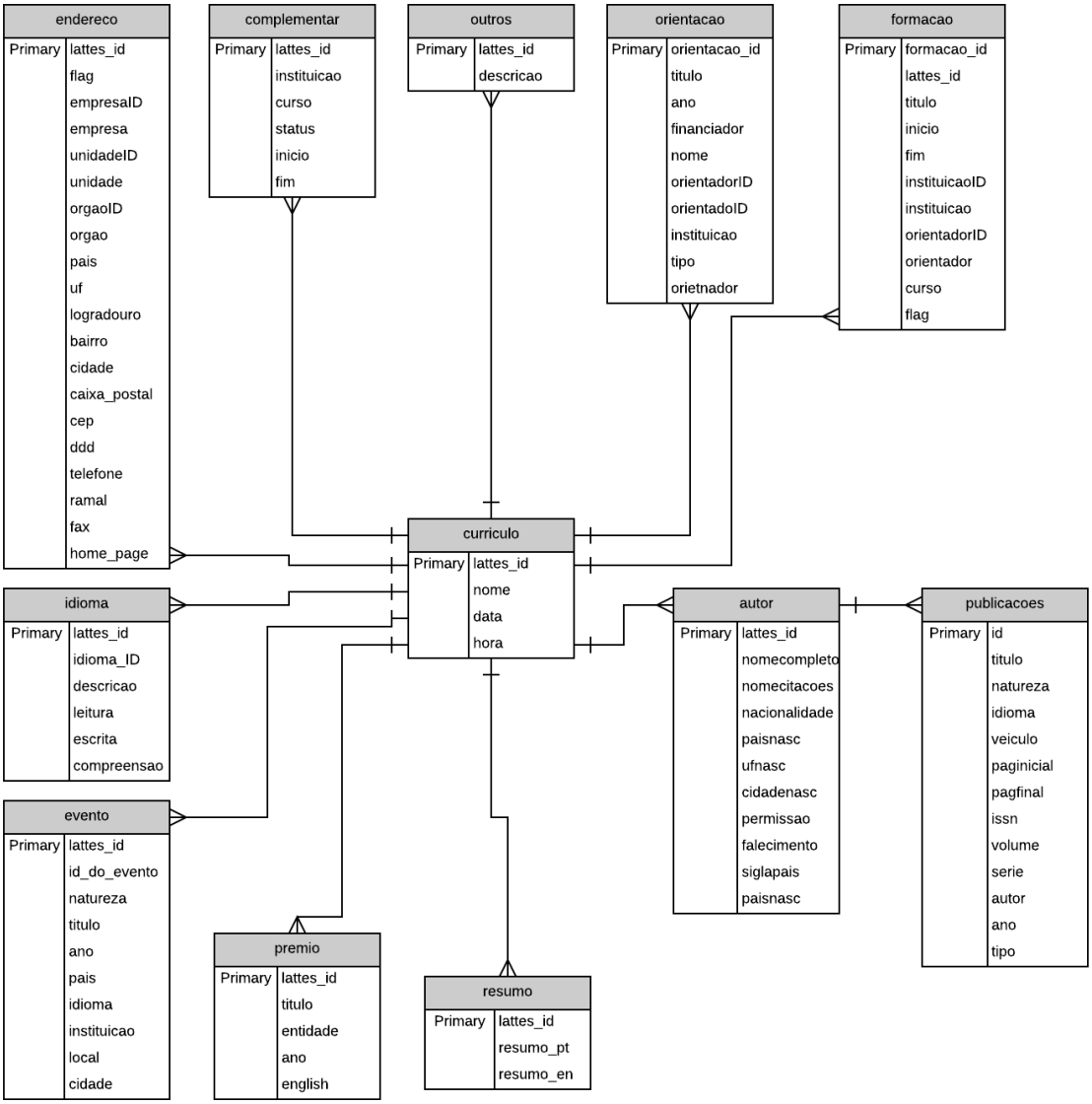
Foi realizado o estudo sobre as tecnologias *PHP, XML e MySQL* para realizar o desenvolvimento da ferramenta/*script* em *PHP* de leitura dos arquivos *XML* e extração automática de informações para armazenar os dados capturados em um banco de dados. Após o estudo técnico das linguagens e ferramentas, e da seleção dos autores mais prolíficos da área de IHC, foi desenvolvido um banco de dados relacional para representar as tabelas que serão criadas no banco de dados *MySQL*. Os atributos expressam exatamente os dados que foram extraídos dos currículos em arquivo *XML*, por isso, alguns dados não necessariamente representam chaves de outras tabelas, como por exemplo, o atributo "instituicaoID"da tabela "formacao". Não há tabelas com as instituições, é apenas uma forma de identificação dentro daquela tabela em específico. O esquema do banco de dados pode ser observado na Figura 4.3. Cada uma das tabelas é descrita a seguir.

- **Currículos:** Tabela contendo as informações de identificação de cada currículo: identificador (lattes_id), nome completo do autor, data e hora da ultima atualização;

- **Formação:** Tabela contendo os dados de formação de cada autor como: título do trabalho, inicio e fim da formação, instituição, orientador, curso;

- **Publicações:** Tabela contendo os dados das publicações como: título da publicação, natureza do publicação, idioma, veículo no qual foi publicado, autores, ano de publicação, tipo de publicação, entre outros;

Figura 4.3 – Diagrama Conceitual do Banco de Dados Desenvolvido.



Fonte: Produção do autor usando o site LucidChart.

- **Autor:** Tabela contendo as informações sobre o autor como: identificador (lattes_id), nome completo do autor, nome usado em citações, nacionalidade, país de nascimento, UF de nascimento, cidade de nascimento, entre outros;

- **Autores_Publicações:** Tabela para vincular cada currículo ao conjunto de publicações cadastradas pelo autor.

- **Orientação:** Tabela contendo as informações de orientações concluídas ou em andamento: título da orientação, ano da orientação, financiador, nome do orientado, nome do orientador, instituição, e tipo de orientação;

- **Resumo CV:** Tabela contendo as informações de resumo de cada autor;

- **Complementar:** Tabela contendo as informações complementares de cada autor: curso realizado, instituição no qual foi realizado, status do curso, inicio e fim do curso;

- **Endereço:** Tabela contendo as informações do endereço de cada autor: logradouro, cidade, UF, país, cep, ddd, telefone, unidade, empresa, entre outros.

- **Idioma:** Tabela contendo os idiomas de cada autor e o nível de instrução do autor para cada aspecto: descrição, leitura, escrita e compreensão.

- **Evento:** Tabela contendo todos eventos de que o autor participou: ID do evento, natureza, título do evento, ano do evento, país, idioma, instituição, local que foi realizada, cidade, entre outros.

- **Prêmio:** Tabela contendo todos os prêmios que o autor recebeu: título do prêmio, entidade, ano do prêmio, entre outros.

- **Outros:** Tabela contendo as outras informações relevantes sobre o autor.

A ferramenta foi implementada usando a linguagem *PHP* com base na extensão *SimpleXML* do *PHP* que permite ler e manipular o *XML*. O *SimpleXML* foi feito para manipular arquivos *XML* que represetam pacotes de dados sobre determinado registro, diferente do *HTML* (PHPMANUAL, 2016). A extensão dispõe de duas funções principais: *simplexml_load_string* e *simplexml_load_file*, onde a primeira carrega um *XML* a partir de uma *string*, enquanto a segunda carrega um *XML* a partir de um nome de um arquivo. As duas funções retornam um objeto da classe *SimpleXMLElement* (classe principal da extensão) (RIBEIRO, 2016).

Na Figura 4.4 podem-se observar uma parte da implementação do *script* usando o *SimpleXML*, onde já é feita a instanciação da classe de conexão ao SGBD *MySQL*, após isso, é realizada a conexão de fato ao banco de dados, e por fim, o código entra em um loop para percorrer todos os itens, que no caso são os currículos dentro da pasta que foram numerados de 1 até 29 com extensão *XML*, por isso o parâmetro '*.xml', para ele percorrer todos os currículos dentro da pasta: *'/var/www/html/curriculos/*.xml'*.

Cada currículo é passado como parâmetro pela função *simplexml_load_file*, e após isso, é definida uma variável em *PHP* para cada uma das informações que serão colocadas na tabela extraindo estas informações com o nome de cada TAG do *XML*,

Figura 4.4 – Código Parcial Implementado em PHP para Extração Automática dos Currículos em XML.

```
<!--*****************************************************************************************************************************-->
<!--* Ferramenta/Script em PHP utilizando o módulo SimpleXML para extrair informações dos currículos da Plataforma Lattes *-->
<!--*****************************************************************************************************************************-->
<!--* Implementação que faz parte do TCC: "Geração das Redes de Colaboração Científica da Comunidade Brasileira de IHC"   *-->
<!--*****************************************************************************************************************************-->
<!--* Autor: Felipe Ciacia de Mendonça                                                                                          *-->
<!--*****************************************************************************************************************************-->

<?php

#Inclue a classe de conexão (criada separadamente):
require_once 'mysqlcon.class.php';

#Instancia a conexão MySQL:
$con = new MySQLCon;

$itens = glob('/var/www/html/curriculos/*.xml');
if ($itens!==false){
    foreach ($itens as $item){
        echo "<br><br>".$item."<br><br>";
        $xml = simplexml_load_file($item);
        #Define as variaveis que serão utilizadas para inserção nas tabelas:
        $id = $xml['NUMERO-IDENTIFICADOR'];
        $nome = $xml->{'DADOS-GERAIS'}['NOME-COMPLETO'];
        $data = $xml['DATA-ATUALIZACAO'];
        $hora = $xml['HORA-ATUALIZACAO'];
        $descricao = $xml->{'DADOS-GERAIS'}->{'OUTRAS_INFORMACOES-RELEVANTES'}['OUTRAS-INFORMACOES-RELEVANTES'];
        $nomecit = $xml->{'DADOS-GERAIS'}['NOME-EM-CITACOES-BIBLIOGRAFICAS'];
        $nacionalidade = $xml->{'DADOS-GERAIS'}['NACIONALIDADE'];
        $paisnasc = $xml->{'DADOS-GERAIS'}['PAIS-DE-NASCIMENTO'];
        $ufnasc = $xml->{'DADOS-GERAIS'}['UF-NASCIMENTO'];
        $cidadenasc = $xml->{'DADOS-GERAIS'}['CIDADE-NASCIMENTO'];
        $permissao = $xml->{'DADOS-GERAIS'}['PERMISSAO-DE-DIVULGACAO'];
        $falecimento = $xml->{'DADOS-GERAIS'}['DATA-FALECIMENTO'];
        $siglapais = $xml->{'DADOS-GERAIS'}['SIGLA-PAIS-NACIONALIDADE'];
        $paisnac = $xml->{'DADOS-GERAIS'}['PAIS-DE-NACIONALIDADE'];
```

Fonte: Produção do autor

podendo ser o elemento de cada TAG, ou até mesmo, os próprios atributos das TAGs. Na Figura 4.5 podem-se ver que após a extração das informações por meio das TAGs do *XML*, é realizada a construção das consultas (*queries*) de inserção no banco de dados, e a inserção no banco com tratamento em caso de erros. A ferramenta criada não necessitará conhecimento prévio do usuário em relação ao arquivo *XML* em que será aplicado a extração, pois mesmo que alguns currículos estejam com menos informações que outros, o código ignorará isto, extraindo apenas os dados existentes no arquivo *XML*.

Figura 4.5 – Código Parcial (2) Implementado em PHP para Extração Automática dos Currículos em XML.

```
        #Query de inserção da tabela CURRICULO:
        $curriculo = "insert into curriculo (lattes_id, nome, data, hora) values ('$id', '$nome', '$data', '$hora')";

        #Query de inserção da tabela OUTROS:
        $other = "insert into outros (lattes_id, descricao) values ('$id', '$descricao')";

        #Query de inserção da tabela AUTORES:
        $autores = "insert into autores (lattes_id, nomecompleto, nomecitacoes, nacionalidade, paisnasc, ufnasc, cidadenasc, permissao, falecimento, siglapais, paisnac) values

        #Query de inserção dos dados da tabela AUTORES_PUBLICACÕES:
        $autores_publicacoes = "insert into autores_publicacoes (lattes_id, publicacaoID, titulo, autor_nome) select lattes_id, id, titulo, nomecompleto from autores INNER JOIN

        #Executa a Query de inserção na tabela CURRICULO:
        $query = mysql_query("$curriculo");

        if ($query){
            echo "Inserção na tabela curriculo realizada com sucesso! ";
            echo "<br>";
        }else{
            echo "Erro na insercao na tabela curriculo!".mysql_errno()." <br><br> ".mysql_error();
            echo "<br>";
        }
```

Fonte: Produção do autor.

Com esta ferramenta, foi possível extrair todas as informações necessárias para as análises dos currículos de todos os autores mais prolíficos de IHC, de acordo com as tabelas desenvolvidas no diagrama da Figura 4.3 e então, popular o banco de dados. Com

o banco de dados relacional pronto, as próximas etapas começaram a ser desenvolvidas que foram o entendimento e preparação dos dados para gerar as tabelas prontas para serem feitas as análises, além da inclusão no banco de dados em grafos gerando assim as redes de colaboração científica.

4.2.3 Entendimento e Preparação dos Dados

Os dados foram inseridos no banco de dados relacional para manter as informações organizadas em tabelas para um posterior estudo, e também para facilitar a filtragem, padronização e limpeza dos dados para que pudessem ser feitas as análises. Além disso, para inserção no banco de dados orientados a grafos Neo4j, as tabelas devem estar em um formato padronizado, ou seja, não era possível extrair os dados diretamente dos arquivos *XML* para o banco de dados em grafos, pois, o Neo4J não suporta este tipo de arquivo diretamente. Desta forma, foi preciso uma conversão para tabelas, na qual foi escolhida a conversão do *XML* para tabelas do banco de dados *MySQL*, e então, exportando estas tabelas para arquivos de extensões *CSV*. Com essas tabelas, é possível tanto já realizar análises bibliométricas e estatistícas com os dados, quanto adaptá-las para serem inseridas no banco de dados *Neo4j*.

Para que os dados fossem devidamente aproveitados, foi necessário realizar uma limpeza e padronização dos dados que foi a realização de correção nas informações ausentes, errôneas ou inconsistentes dentro das tabelas, e além disso, padronizar todos os nomes dos autores, nomes das instituições, nomes dos veículos de publicações, processo que demandou bastante tempo por ser um processo semi-manual. Esta etapa foi um dos grandes problemas encontrados, pois foram encontrados diversos problemas de de padronização de nomes, tais como os nomes dos eventos, pessoas e instituições. Um exemplo foi o próprio Simpósio IHC: “Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais”. Foram encontrados mais de 20 nomes diferentes, como pode ser visto alguns deles na Figura 4.6, causado pela mudança do nome do evento durante os anos, erros de escrita dos autores, ou até mesmo por causa da edição do evento que acompanhava o nome. Este processo de padronização e limpeza dos dados foi iniciada em maio de 2017, e finalizada em meados de junho de 2017.

Figura 4.6 – Exemplo de nomes não padronizados encontrados

IHC 2015
IHC'99 II Workshop sobre fatores Humanos em Sistemas Computacionais
IHC 2013
VIII Simpósio Brasileiro de Fatores Humanos em Sistemas Computacionais (IHC 2008)
XI Simpósio de Fatores Humanos em Sistemas Computacionais, IHC
V Symposium on Human Factors in Computer Systems (IHC2002)
IHC 2011 - X Simpósio Brasileiro de Fatores Humanos em Sistemas Computacionais
X Simpósio Brasileiro de Fatores Humanos em Sistemas Computacionais (IHC + CLIHC 2011)
IHC '12 - 11th Brazilian Symposium on Human Factors in Computing System

Fonte: Produção do autor.

4.2.4 Transformação dos Dados

Após a fase de pré-processamento dos dados, no qual foi realizado a limpeza, padronização e filtragem dos dados, as tabelas já estavam aptas para realizar as análises bibliométricas e estatísticas, porém, algumas destas tabelas ainda não estavam no formato adequado para a inserção no banco de dados Neo4j, isso porque alguns requisitos são necessários para criar alguns relacionamentos dentro das *querys* de consulta *Cypher* (Linguagem de pesquisa em grafos utilizada pelo Neo4j) que serão explicados a seguir. Na construção da rede de coautoria formada pelos pesquisadores e todos os seus coautores por exemplo, a tabela como pode ser vista na Figura 4.7 continha em cada linha o nome do título da publicação, e ao lado o nome dos autores daquela publicação incluindo os autores prolíficos. Porém, para criar as conexões entre os nós dentro da rede no *Neo4J*, na *query* foi preciso especificar qual coluna será o "Coautor" que será ligado com o "Pesquisador" que teria que estar em outra coluna, que não existia na tabela original.

Figura 4.7 – Trecho da tabela de publicações antes da Transformação

5Cam: a multicamera system for panoramic capture of videos	Camilo Telles Pereira Santos	TRABALHO EM EVENTO
5Cam: a multicamera system for panoramic capture of videos	Celso Alberto Saibel Santos	TRABALHO EM EVENTO
7x1PT: um Corpus extraído do Twitter para Análise de Sentimentos em Língua Portuguesa	Isabel Harb Manssour	TRABALHO EM EVENTO
7x1PT: um Corpus extraído do Twitter para Análise de Sentimentos em Língua Portuguesa	Milene Selbach Silveira	TRABALHO EM EVENTO
7x1PT: um Corpus extraído do Twitter para Análise de Sentimentos em Língua Portuguesa	Sílvia Maria Wanderley Moraes	TRABALHO EM EVENTO

Fonte: Produção do autor.

Desta forma, foi necessário implementar um código em *Python* representado na Figura 4.8 que recebesse como entrada a tabela de publicações, e retornasse como saída uma nova tabela com as mesmas publicações, porém com uma nova coluna que especificasse qual daqueles autores mais prolíficos realizaram a coautoria com aqueles autores. O código foi escrito para funcionar da seguinte forma:

- Ler tabela de publicações;
- Identificar o título da publicação;

- Varrer a tabela em busca de publicações com o mesmo título;
- Quando achar a publicação de mesmo título, verificar se o autor é um dos autores prolíficos (Guardados em uma lista no código);
- Se autor é um dos autores prolíficos, guardar em uma nova lista, senão segue em diante;
- Quando não houver mais publicações com o mesmo título, o código irá varrer todas as publicações de novo procurando novamente publicações com o mesmo título;
- Assim que achar as publicações com o mesmo título, o código escreve em uma nova tabela estas mesmas linhas contendo esta publicação porém incluindo uma coluna com o(s) nome(s) do autor(es) prolífico(s) (guardados em uma lista anteriormente) no qual aquele autor contém um relacionamento.

Figura 4.8 – Trecho do código em Python para Tranformação de Tabelas

```
#ler arquivo csv

ofile = open('novo.csv', "wb")
fieldnames = ['TITULO', 'NATUREZA', 'IDIOMA', 'VEICULO', 'INTERNACIONALIDADE', 'N_AUTOR', 'NOME', 'ANO', 'TIPO', 'PESQUISADOR']
writer = csv.DictWriter(ofile, fieldnames=fieldnames)

with open('publicacoes_nova.csv') as csvfile:
        reader = csv.DictReader(csvfile)
        for row in reader:
                if (row not in lista):
                        lista.append(row);
                        titulo = row['TITULO']
                        print ('1) FOR', titulo)
                        if (titulo != titulo_antigo):
                                titulo_antigo = titulo;
                                csvfile.seek(0)
                                for row in reader:
                                        #print ('2) FOR')
                                        if (titulo == row['TITULO']):
                                                print ('2-1) FOR')
                                                if (row['AUTORES'] in autores):
                                                        print ('2-2) FOR')
                                                        autor = row['AUTORES']
                                                        autores2.append(autor)

                                tamanho_lista = len(autores2)
                                print ("tamanho da lista:", tamanho_lista)
                                csvfile.seek(0)
                                for x in range (tamanho_lista):
                                        print ('3) FOR')
                                        for row in reader:
                                                if (titulo == row['TITULO']):
                                                        print ('4) FOR')
                                                        writer.writerow({'TITULO':row['TITULO'], 'NATUREZA':row['NATUREZA'], 'IDIOMA':row
['IDIOMA'], 'VEICULO': row['VEICULO'], 'INTERNACIONALIDADE': row['INTERNACIONAL'], 'N_AUTOR': row['N_AUTOR'], 'NOME': row['AUTORES'], 'ANO':
row['ANO'], 'TIPO': row['TIPO'], 'PESQUISADOR': autores2[x]})
                                                        csvfile.seek(0)
                                autores2[:] = []
```

Fonte: Produção do autor.

Uma parte da tabela gerada após a Transformação é mostrada na Figura 4.9. Ela contém todos os relacionamentos, os quais possuem uma coluna dos autores da publicação e ao lado, o autor prolífico no qual os autores das publicações foram relacionados.

Um problema na geração desta tabela foi a duplicidade de informações, como por exemplo, um autor que foi ligado a ele mesmo, ou um dos autores prolíficos que foi ligado a um outro autor prolífico diferente (dentro da lista dos 29 autores mais prolíficos)

Figura 4.9 – Trecho da tabela publicações após a Transformação

5Cam: a multicamera system for panoramic capture of videos	Camilo Telles Pereira Santos	TRABALHO EM EVENTO	Celso Alberto Saibel Santos
5Cam: a multicamera system for panoramic capture of videos	Celso Alberto Saibel Santos	TRABALHO EM EVENTO	Celso Alberto Saibel Santos
7x1PT: um Corpus extraído do Twitter para Análise de Sentimentos em Língua Portuguesa	Isabel Harb Manssour	TRABALHO EM EVENTO	Milene Selbach Silveira
7x1PT: um Corpus extraído do Twitter para Análise de Sentimentos em Língua Portuguesa	Milene Selbach Silveira	TRABALHO EM EVENTO	Milene Selbach Silveira
7x1PT: um Corpus extraído do Twitter para Análise de Sentimentos em Língua Portuguesa	Silvia Maria Wanderley Moraes	TRABALHO EM EVENTO	Milene Selbach Silveira

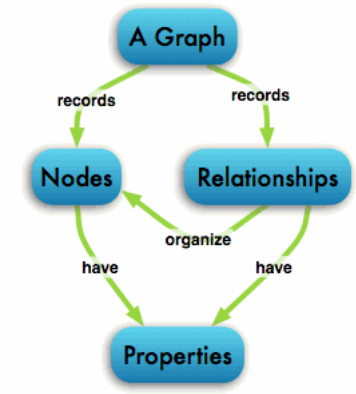
Fonte: Produção do autor.

mais de uma vez, porém isto não foi uma complicação porque o próprio Neo4J na hora de inserir no banco de dados contém a premissa de manter sempre apenas um relacionamento, além do que também é possível especificar se quer ou não relacionamentos entre o próprio nó. Então todos estes cuidados foram tomados, e as redes foram geradas devidamente.

4.2.5 Geração das Redes de Colaboração Científica

Após a última fase em que os dados foram preparados, além de ser possível realizar as primeiras análises, também foi possível realizar a inserção dos dados no banco de dados orientado a grafos *Neo4J*. O *Neo4J* é um banco de dados orientado a grafos, no qual é preciso criar nós que seriam equivalentes as entidades de um banco de dados relacional, e posteriormente, os relacionamentos entre estes nós, que seriam os próprios relacionamentos de um banco de dados relacional convencional. Porém, o diferencial do banco de dados *Neo4J* é sua representação em forma de grafos, o que torna o processamento muito mais rápido, e menos custoso, além de deixar a organização dos dados mais simples, não necessitando de diversas tabelas como um banco de dados relacional.

Figura 4.10 – Representação do banco Neo4J



Fonte: http://neo4j.com/developer/graph-database/

Na Figura 4.10 pode-se ver a representação de como funciona os dados no banco *Neo4J* onde existe um grafo que grava nós e relacionamentos, tais que organizam estes nós, e além disso, tanto os nós quanto os relacionamentos armazenam atributos.

Tabela 4.1 – Tipos de nós e suas respectivas quantidades inseridas no Neo4J

Tipo de Nó	**Descrição**	**Quantidade**
Pesquisador	Autores mais prolíficos de IHC	29
Orientado	Orientados de Mestrado e/ou Doutorado	624
Coautor	Coautores dos pesquisadores	2316
Veículo	Veículos de Publicação	1199
Instituição	Universidades as quais os autores são vinculados	43

Tabela 4.2 – Tipos de Relacionamentos e suas respectivas quantidades inseridas no Neo4J

Tipo de Relacionamento	**Descrição**	**Quantidade**
Doutorado	Orientações de Doutorado	153
Mestrado	Orientações de Mestrado	547
Coautoria	Coautoria entre os autores das publicações	2949
Publicou	Conexão entre os pesquisadores e os veículos que ele publicou	2093
Vínculo	Vínculo entre os pesquisadores e a universidade no qual ele é vinculado	659

Para a geração da rede, foram reunidas as tabelas no armazenamento do banco *Neo4J* e realizado a importação das tabelas e posteriormente, a criação dos nós e relacionamentos da rede de colaboração científica.

Para inserção dos nós, foram realizadas diversas *querys* como o exemplo mostrado na Figura 4.11 na qual está inserindo todos os nós "Pesquisadores"e seus respectivos atributos contidos na tabela. Os nós foram divididos em 5 tipos como segue na Tabela 4.1 com suas respectivas descrições e quantidades inseridas no banco *Neo4J*.

Figura 4.11 – *Query* de inserção dos nós "Pesquisadores"no banco Neo4J

```
LOAD CSV WITH HEADERS FROM "file:///pesquisadores.csv" AS row
CREATE (:Pesquisador {pesquisadorID: row.LATTESID, Nome: row.NOME, Pais: row.PAIS,  Estado: row.ESTADO,
Cidade: row.CIDADE, Data_atualizacao: row.DATA, Hora_atualizacao: row.HORA, Grau_instrucao: row.GRAU,
Instituicao: row.UNIV, Endereco: row.ENDERECO, Idioma_ingles: row.INGLES, Idioma_espanhol: row.ESPANHOL,
Idioma_portugues: row.PORTUGUES, Idioma_frances: row.FRANCES, Idioma_italiano: row.ITALIANO,  Idioma_alemao:
row.ALEMAO, Idioma_libras: row.LIBRAS, Idioma_japones: row.JAPONES, N_Papers: row.PAPERS});
```

Fonte: Produção do autor.

Após a inserção dos nós, foi realizada a criação das conexões entre os nós, foi necessário também importar as tabelas para que o banco conseguisse criar os relacionamentos por meio de *querys* como no exemplo da Figura 4.12 em que é criado o relacionamento entre os "Pesquisadores"e os "Veículos de Publicação". Os relacionamentos foram divididos em 5 tipos também, assim como segue na Tabela 4.2 com suas descrições e quantidades inseridas.

Figura 4.12 – *Query* de inserção do relacionamento "Vinculo"entre "Pesquisadores"e "Veiculos"no banco Neo4J

```
USING PERIODIC COMMIT
LOAD CSV WITH HEADERS FROM "file:///relacionamentos_pesquisadores_veiculos.csv" AS row
MATCH (v:Veiculo{Veiculo: row.VEICULO})
MATCH (p:Pesquisador{Nome: row.AUTORES})
MERGE (p)-[c:Publicou]->(v)
```

Fonte: Produção do autor.

Desta forma, foram criadas 5 sub-redes (grafos) no banco em que todas elas estão inseridas pelo menos os 29 pesquisadores prolíficos ilustrados na Figura 4.13, formando assim, uma Rede de Colaboração Científica da Comunidade Brasileira de IHC.

Figura 4.13 – Grafo de Ilustração dos Tipos de Nós e Relacionamentos gerados no Neo4J

Fonte: Produção do autor.

4.3 Análise dos Resultados

Com base nos dados extraídos dos currículos dos autores mais prolíficos de IHC, foram realizadas as análises bibliométricas e estatísticas. Entre os resultados tivemos várias avaliações sobre a comunidade brasileira de IHC, entre elas, como está a internacionalização da comunidade em relação as suas publicações, quais veículos os pesquisadores da comunidade tem publicado e as suas respectivas qualidades, como o trabalho realizado pelos pesquisadores tem se difundido além das fronteiras do fórum nacional de IHC, entre outras análises sobre a comunidade brasileira de IHC envolvendo tanto o Simpósio de IHC, como todos os veículos nos quais os autores prolíficos publicaram.

Partindo das redes geradas no *Neo4J*, foi possível obter algumas métricas em

relação as redes sociais formadas, como diversas medidas de centralidade que determinam a importância de um nó dentro do grafo, ou seja, no caso das redes geradas neste trabalho, determinamos os nós (Pesquisadores) centrais ou mais influentes dentro desta rede social. Além das centralidades, também foi possível verificar a formação de grupos ou comunidades dentro da rede social, com o objetivo de definir grupos de nós que possuem mais conexões entre si do que com o resto da rede social (FORTUNATO, 2010). Tomando como base a medida do Número de Erdös apresentado no Capítulo 2, podemos adaptá-la e formar uma medida de distância de qualquer autor da comunidade brasileira de IHC ao autor mais influente da comunidade.

4.3.1 Análises Bibliométricas e Estatísticas

É importante salientar que o período de coleta de dados foi de dezembro de 2015 a setembro de 2016, iniciando pela coleta de todas edições dos anais do Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (IHC). Os anais analisados são do perído desde 1998 (primeira edição do IHC) até o ano de 2015 quando iniciou-se a coleta de dados. De um total de 340 artigos completos, foram verificados quais autores mais publicaram, e então selecionado todos os autores que publicaram cinco ou mais artigos completos. A escolha de cinco ou mais artigos está relacionada com as etapas seguintes do trabalho, pois, caso a seleção de autores fosse autores que produziram menos artigos, o número de currículos aumentaria consideravelmente, e tornaria cada vez mais inviável de realizar a captura. Como a coleta dos currículos dos autores que seria feita posteriormente teria que ser manual, pois a Plataforma Lattes não disponibilizava na época uma ferramenta automática para extração, foi escolhido o número de cinco ou mais publicações, pois assim, foi tido como resultado um total de 29 autores mais prolíficos, mostrado na Tabela 4.3, número regular para se extrair os dados, e que resultaria em um número de informações suficientes para realizar esta pesquisa.

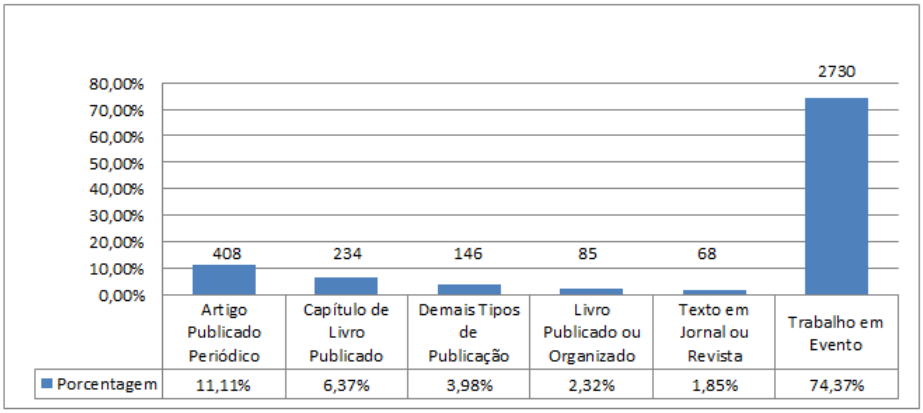
Desse total de 340 artigos completos, 242 destes artigos (71%) tiveram participação de pelo menos um dos 29 autores prolíficos, o que mostra que esses 29 autores realmente são expressivos na comunidade brasileira de IHC. Partindo dos currículos destes autores, foram encontrados 3671 publicações ao total, sendo elas de vários tipos de publicações de acordo com a classificação do currículo da Plataforma Lattes, entre elas: Artigos Publicados em Periódicos (o que comumente chamamos de *Journals*), Trabalhos

Tabela 4.3 – Os 29 autores mais prolíficos do IHC

Autores	**Autores**
Artur Henrique Kronbauer	Marcelo Soares Pimenta
Carla Faria Leitão	Marco Antônio Alba Winckler
Carla Maria Dal Sasso Freitas	Maria Cecília Calani Baranauskas
Celso Alberto Saibel Santos	Maria Elizabeth Sucupira Furtado
Clarisse Sieckenius de Souza	Milene Selbach Silveira
Cristiano Maciel	Raquel Oliveira Prates
Denis Silva da Silveira	Roberto Pereira
Heloísa Vieira da Rocha	Sérgio Roberto Pereira da Silva
Isabela Gasparini	Simone Bacellar Leal Ferreira
Jair Cavalcanti Leite	Simone Diniz Junqueira Barbosa
Janne Y. Y. Oeiras Lachi	Tayana Uchôa Conte
Junia Coutinho Anacleto	Vânia Paula de Almeida Neris
Lara S. Godoy Piccolo	Vinícius Carvalho Pereira
Leonardo Cunha de Miranda	Walter de Abreu Cybis
Lucia Vilela Leite Filgueiras	

em Eventos (podendo ser um Artigo Completo, Artigo Resumido ou Resumo Expandido), Capítulos de Livros Publicados, Livros Publicados ou Organizados, Textos em Jornal ou Revista (do tipo *Magazine*) e Demais Tipos de Publicações. A publicação mais antiga encontrada foi do ano de 1981 e o ano mais recente até a data da coleta foi o ano de 2016, porém existiam muitas publicações sem o ano de publicação.

Figura 4.14 – Tipos de Publicação

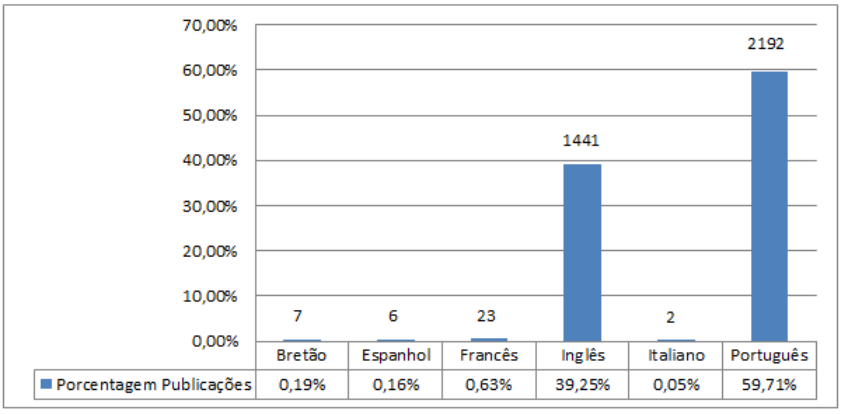


Fonte: Produção do autor.

Como pode ser visto na Figura 4.14, as publicações em Trabalhos em Eventos (74,37%) e as publicações em *Journals* (11,11%) representam mais e 85% de todas as publicações, sendo os outros tipos com quantidades insuficientes para retirar qualquer análise concisa, por este motivo, restringimos a maioria das análises a estes tipos de publicações.

Já em relação aos idiomas encontrados nas publicações, tivemos 59,71% dos trabalhos publicados em Português e 39,25% em Inglês como mostrado na Figura 4.15, o que mostra a importância de trabalhos publicados em Inglês, mesmo que ainda seja um número menor que o Português. Ainda tivemos trabalhos encontrados nos idiomas Espanhol, Francês, Italiano, e até mesmo no dialeto Bretão.

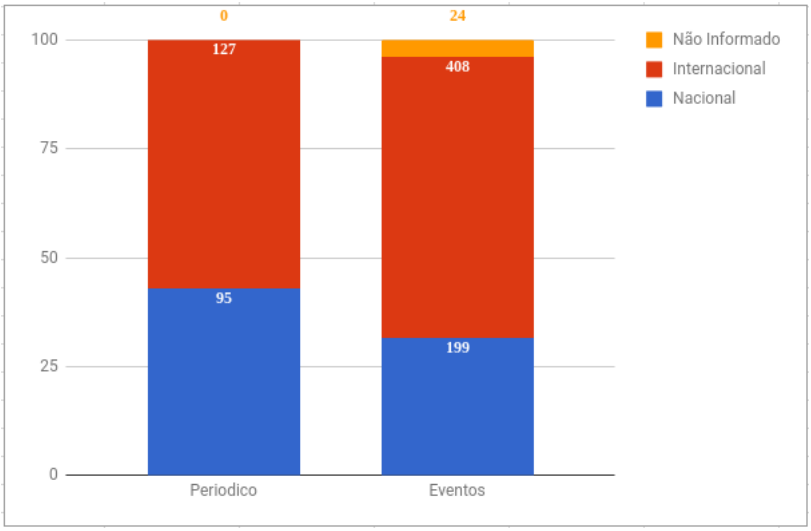
Figura 4.15 – Total de Publicações por Idioma



Fonte: Produção do autor.

Em relação a nacionalidade das publicações como mostrado na Figura 4.16, ou seja, a nacionalidade dos veículos em que foram publicados, separados por Nacionais e Internacionais, foi observado a preocupação da comunidade em internacionalizar seus trabalhos e pesquisa já que 62% das publicações foram em Veículos de Publicação Internacionais, enquanto 34% foram em Veículos de Publicação Nacionais, e outros 4% dos veículos não foram informados, ou não encontrados.

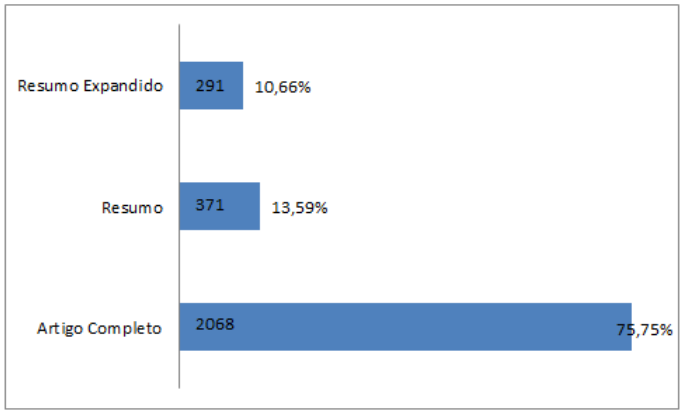
Figura 4.16 – Veículos de Publicações Nacionais e Internacionais



Fonte: Produção do autor.

Verificando apenas as publicações em Trabalhos em Eventos, elas eram subdivididas em Artigos Completos, Resumo, e Resumo Expandidos como ilustrado no gráfico da Figura 4.17, onde mais de 75% dos trabalhos são Artigos Completos, o que mostra a preocupação em publicar resultados finais de pesquisas realizadas e trabalhos mais completos.

Figura 4.17 – Publicações em Trabalhos em Eventos



Fonte: Produção do autor.

Seguindo a evidenciação dos periódicos (*Journals*) e eventos os quais os autores prolíficos do IHC publicaram, foi feito um ranqueamento dos veículos de publicações nos quais tiveram mais publicações, divididos pelo tipo de publicação, e também pela nacionalidade considerando novamente, publicações nacionais e internacionais. É importante salientar que cada publicação, mesmo sendo coautoria, é contada uma única vez. Também foi capturado o Qualis (Conjunto de procedimentos utilizado pela Capes para classificação da qualidade da produção intelectual dos programas de pós-graduação) de cada publicação para que pudesse analisar qual a qualidade dos veículos de publicações os autores têm publicado.

Com base nos periódicos internacionais nos quais tiveram mais publicações dos pesquisadores de IHC foi realizado o ranqueamento como mostrado na Tabela 4.4. Como muitos veículos ficaram empatados na 6ª posição, optou-se por apresentar a tabela até a posição 6 (*Top 6*).

Partindo da mesma premissa da última análise, foi realizado o ranqueamento dos periódicos nacionais. Como pode ser visto na 4.5, o ranqueamento foi feito até o *Top* 10. Entre os veículos mais utilizados, destaca-se alguns que são da área de Informática na Educação, demonstrando a interdisciplinaridade dos autores das publicações assim como a própria área de IHC. No que se refere a periódicos nacionais, é importante ressaltar que

Tabela 4.4 – "Top 6"Periódicos Internacionais

Pos.	Nome	Qtde	Qualis
1	Interactions (New York, N. Y.)	11	B1
2	Interacting with Computers	8	A2
3	Procedia Computer Science	6	C
4	Revista IEEE América Latina	5	B4
	Universal Acess in the Information		B1
5	CLEI Electronic Journal	4	B5
	Educational Technology & Society		B4
	Human-Computer Interaction		-
6	Computer in Entertainment	3	B1
	Information Polity		B5
	International Journal for Infonomics (IJI)		C
	International Journal of Continuing Engineering Education		-
	International Journal of Human-Computer Studies		A1
	Journal of Information and Data Management - JIDM		B3
	Journal of Systems and Software		A2
	Knowledge-Based Systems		A1
	Lecture Notes in Computer Science		C
	SIGCHI Bulletin		-

o periódico que é o *JIS* (*SBC Journal on Interactive Systems*) tem foco na área, enquanto outros apresentados na tabela são mais gerais.

Na Tabela 4.6 são apresentados os "*Top* 10"eventos internacionais com maior número de publicações dos autores, ressaltando que podem ser artigos completos, resumos ou resumos expandidos. No entanto, na Plataforma Lattes, se não estiver especificado que a publicação é um *workshop* relacionado a um evento principal, este evento irá se destacar, pois acaba contando como se tivesse publicações de artigos como no caso do CHI, 6ª colocada na tabela 4.6, em que pesquisadores publicaram não necessariamente na

Tabela 4.5 – "Top 10"Periódicos Nacionais

Pos.	Nome	Qtde	Qualis
1	Revista Brasileira de Informática na Educação	19	B3
2	Revista Novas Tecnologias na Educação (RENOTE)	18	B5
3	Journal of the Brazilian Computer Society	14	B1
4	Cadernos de Informática (UFRGS)	12	-
5	Revista de Informática Teórica e Aplicada	10	B3
6	SBC Journal on 3D Interactive Systems	9	B3
7	Revista Brasileira de Administração Científica	7	B5
8	Cadernos do IME. Série Informática	6	B5
9	iSys: Revista Brasileira de Sistemas de Informação	5	B3
10	Mergulhar A Descoberta do Mar	4	-

Tabela 4.6 – "Top 10"Eventos Internacionais

Pos.	Nome	Qtde	Qualis
1	International Conference on Human-Computer Interaction (HCI)	126	B2
2	Congresso Latino-Americano de Interação Humano-Computador (CLIHC)	75	B4
3	International Conference on Enterprise Information System (ICEIS)	63	B2
4	IFIP TC International Conference on Human-Computer Interaction (INTERACT)	52	A2
5	Conferencia Latinoamericana en Informática (CLEI)	33	B3
6	ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)	30	A1
7	IADIS International Conference WWW/Internet (ICWI)	28	-
8	International Association for Management of Techonology (IAMOT)	21	-
9	International Conference on Software Engineering & Knowledge Engineering (SEKE)	18	B1
10	Conferência do Conselho Latino Americano das Escolas de Administração (CLADEA)	17	-
	International Conference on Informatics and Semiotics in Organizations (ICISO)		B2

Tabela 4.7 – "Top 10"Eventos Nacionais

Pos.	Nome	Qtde	Qualis
1	Simpósio Brasileiro sobre Fatores Humanos em Sistemas Co mputacionais (IHC)	384	B2
2	Simpósio Brasileiro de Informática na Educação (SBIE)	107	B1
3	Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)	60	B2
4	Congresso da Sociedade Brasileira de Computação (CSBC)	53	-
5	Escola Regional de Informática da SBC (ERI)	50	-
6	Simpósio Brasileiro de Computação Gráfica e Processamento de Imagens (SIBGRAPI)	39	B1
7	Simpósio Brasileiro de Engenharia de Software (SBES)	32	B2
8	Encontro Nacional dos Programas de Pós Graduação e Pesquisa em Administração (ENANPAD)	23	-
9	Simpósio Brasileiro de Sistemas de Informação (SBSI)	22	B2
10	Encontro de Atividades Científicas da UNOPAR	21	-
	Simpósio Brasileiro de Qualidade de Software (SBQS)		B3

trilha principal.

É possível ver que as duas primeiras colocadas são especificamente da área de IHC, assim como outras grandes conferências da área também fazem parte do alvo de veículos de publicação dos autores (inclusive alguns destes autores fazem parte do corpo de revisão de algumas destas conferências). Outras conferências (como o *ICEIS* e o *ICWI*, por exemplo) não são eventos da área, porém, contém trilhas relacionadas a ela.

A Tabela 4.7 apresenta os "*Top* 10"eventos nacionais, com destaque para o próprio Simpósio IHC com 384 publicações. Pode-ser observar diversas conferências e simpósios ligados à Sociedade Brasileira de Computação (SBC) apresentados em detalhe na Tabela 4.8 que mostra todos eventos ligados à SBC em que os autores já publicaram (em ordem alfabética, incluindo o IHC).

Após estas análises de quais foram os veículos com maiores publicações, uma nova questão surgiu: **Como os pesquisadores estão inserindo suas pesquisas em veículos de qualidade na comunidade internacional?**. Partindo dos eventos e periódicos de maior visibilidade segundo o Google Scholar que usa um índice chamado índice-h5 criamos o ranqueamento "*Top* 20"para eventos ou periódicos da área de IHC como podemos ver na Tabela 4.9 no qual indicar a quantidade de publicações dos autores

Tabela 4.8 – Eventos ligados à SBC em que os autores do IHC já publicaram artigos

Nome dos Eventos Ligados a SBC
Congresso Brasileiro de Informática na Educação (CBIE)
Congresso Brasileiro de Software (CBSOFT)
Congresso da Sociedade Brasileira de Computação (CSBC)
Seminário Integrado de Software e Hardware (SEMISH)
Simpósio Brasileiro de Arquitetura e Computadores - Processamento de Alto Desempenho
Simpósio Brasileiro de Banco de Dados (SBBD)
Simpósio Brasileiro de Componentes, Arquitetura e Reutilização de Software (SBCARS)
Simpósio Brasileiro de Computação Gráfica e Processamento de Imagens (SIBGRAPI)
Simpósio Brasileiro de Computação Musical (SBCM)
Simpósio Brasileiro de Computação Ubíqua e Pervasiva (SBCUP)
Simpósio Brasileiro de Engenharia de Software (SBES)
Simpósio Brasileiro de Games e Entretenimento Digital (SBGames)
Simpósio Brasileiro de Informática na Educação (SBIE)
Simpósio Brasileiro de Inteligência Artificial (SBIA)
Simpósio Brasileiro de Linguagens de Programação (SBLP)
Simpósio Brasileiro de Qualidade de Software (SBQS)
Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)
Simpósio Brasileiro de Segurança da Informação e Sistemas Computacionais (SBSeg)
Simpósio Brasileiro de Sistemas Colaborativos (SBSC)
Simpósio Brasileiro de Sistemas de Informação (SBSI)
Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)
Simpósio Brasileiro de Tecnologia de Informação e Linguagem Humana (STIL)
Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (IHC)
Symposium on Virtual and Augmented Reality (SVR)
Workshop de Computação Aplicada em Governo Eletrônico (WCGE)
Workshop de Desafios da Computação Aplicada à Educação (DesafIE)
Workshop de Informática na Escola (WIE)
Workshop de Realidade Virtual e Aumentada (WRVA)
Workshop de TV Digital e Interativa (WTVDI)
Workshop sobre Educação em Computação (WEI)

Tabela 4.9 – Veículos do "*Top* 20"do Google Scholar na área de IHC e a quantidade de publicações dos autores

Pos.	Nome Evento/Periódico	Qtde
1	Computer Human Interaction (CHI)	30
2	ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)	5
3	ACM Symposium on User Interface Software and Technology (UIST)	-
4	ACM Conference on Pervasive and Ubiquitous Computing (UbiComp)	-
5	IEEE Transactions on Affective Computing	-
6	ACM/IEEE International Conference on Human Robot Interaction	-
7	International Journal of Human-Computer Studies	3
8	Mobile HCI	1
9	ACM Transactions on Computer-Human Interaction (TOCHI)	1
10	Behaviour & Information Technology	-
11	Interacting with Computers	8
12	International Conference on Affective Computing and Intelligent Interaction and Workshops	-
13	International Conference on Multimodal Interfaces (ICMI)	1
14	IEEE International Symposium on Mixed and Augmented Reality	-
15	International Journal of Human-Computer Interaction	-
16	International Conference on Intelligent UserInterfaces (IUI)	3
17	IFIP Conference on Human-Computer Interaction (INTERACT)	52
18	International Conference on Tangible, Embedded and Embodied Interaction	-
19	Conference on Designing Interactive Systems	2
20	IEEE Transactions on Haptics	-

em cada um destes *Top* veículos de publicação.

O índice-h (Proposta para quantificar a produtividade e o impacto de pesquisadores com base nos seus artigos mais citados (HIRSCH, 2007)) é utilizado pelo Google Scholar, calculando usando apenas os artigos publicados nos últimos 5 anos completos recentes. Simplificando, o índice-h é o número de artigos com citações maiores ou iguais a esse número. Por exemplo: um pesquisador com h = 6 tem 6 artigos que receberam 6 ou mais citações; Assim como um departamento de computação com h = 90 tem 90 artigos com 90 ou mais citações e assim por diante. Porém o índice-h5 do Google Scholar utiliza o índice-h apenas nos últimos 5 anos completos recentes (SCHOLAR, 2017).

Entras as análises que envolviam as Leis clássicas de Bibliometria citadas no Capítulo 2, uma análise interessante seria a de Obsolência ou Meia vida das publicações, porém como aqui não trabalhamos com as citações/referências bibliográficas das publicações, esta análise não foi considerada adequada para este trabalho, assim como não foi possível realizar a Lei do Elitismo que também necessitava das citações das publicações.

Também não foi realizada a Lei de Zipf, porque esta lei é uma análise com foco local em um único documento ou um conjunto de documentos específicos, já que objetiva buscar as palavras mais encontradas no(s) artigo(s). E nesta pesquisa não tivemos acesso a todas as publicações em si dos autores prolíficos, e sim, somente a quais são estas

Tabela 4.10 – *Top 7* das palavras mais encontradas nos títulos das publicações

Palavra-Chave	**Quantidade**
Approach	123 (2%)
Based	115 (2%)
Framework	100 (2%)
Design	84 (1%)
User	79 (1%)
Systems	74 (1%)
Software	69 (1%)

publicações. O mesmo sucedeu-se para a Lei de Goffman que é uma variante da Lei de Zipf.

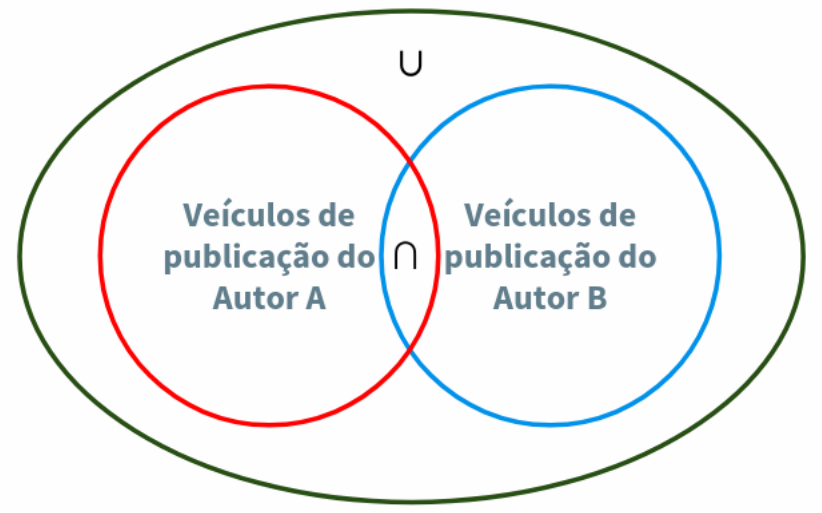
O que foi possível realizar foi a busca das palavras mais encontradas nos títulos de todas as publicações dos autores prolíficos de IHC resultando em um *Top 7* de palavras como podemos ver na Tabela 4.10. Não é uma análise tão efetiva caso fosse feito com os textos das publicações de fato e um detalhe que pode ser verificado é que todas as palavras mais encontradas estão na língua inglesa.

Similaridade de Jaccard

No que se refere a similaridade entre os autores, ou seja, verificar se os autores publicam em veículos de publicações em comum, podendo analisar se possuem perfis similares como pode ser visto na Figura 4.18 que foi produzida usando a ferramenta Plotly. Uma forma de quantificar a similaridade entre comunidades ou indivíduos é usar coeficientes de similaridade, entre eles se destaca o coeficiente de similaridade de *Jaccard* (TAN; STEINBACH; KUMAR, 2009). Este índice ou coeficiente de similaridade compara dois conjuntos de dados, os quais contém dados iguais e dados distintos um do outro, e então, verifica quais destes dados eles têm em comum em comparação aos dados que tem distintos. E com isto temos um valor dentro do intervalo que vai de 0 a 1, sendo 1 o mais similar possível, e 0 totalmente diferente. Com base nisso, foram elaboradas as similaridades dos autores mais prolíficos analisando em pares individualmente para verificar o quão cada par de autores era semelhante.

O gráfico representado na Figura 4.19 representa os pares de autores, onde cada pequeno espaço na grade inferior é um autor, assim como cada pequeno espaço do lado direito da grade, formando dessa forma os pares de autores. O intervalo do gráfico vai de 0 a 1, sendo assim, no mapa de calor, onde está mais escuro indica que aquele par está

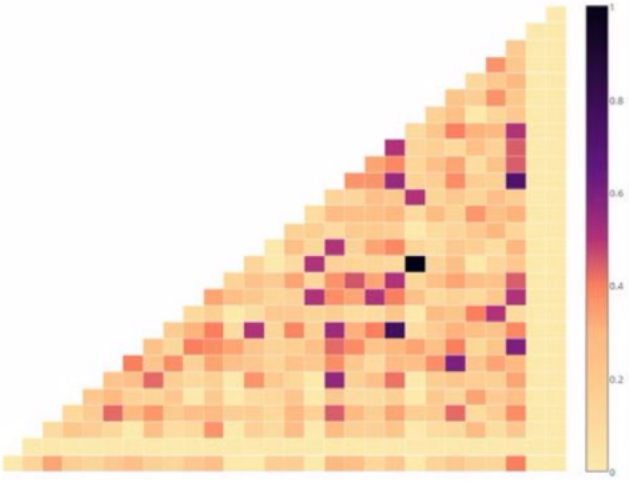
Figura 4.18 – Representação dos conjuntos analisados para a Similaridade de Jaccard



Fonte: Produção do autor.

mais similar, enquanto os pares mais claros indica que ele está menos similar. Inicialmente foram elaborados as similaridades dos autores para todos os tipos de publicações, porém, os números de similaridade ficaram muito baixos, a ponto de quase não ser visível num mapa de calor, pois os valores ficavam muito próximos de zero (0). Então, para que a visualização se tornasse mais efetiva, utilizamos algumas restrições até que o mapa de calor pudesse se tornar mais visivelmente efetivo. Primeiro foram restringidos os artigos apenas de eventos, posteriormente só eventos nacionais, depois só eventos nacionais ligados a SBC, e finalmente eventos ligados a SBC sem o IHC (pois todos os autores já contém o IHC). Dessa forma os valores encontrados se tornaram bem mais vísiveis no mapa de calor como visto na Figura 4.19 anteriormente.

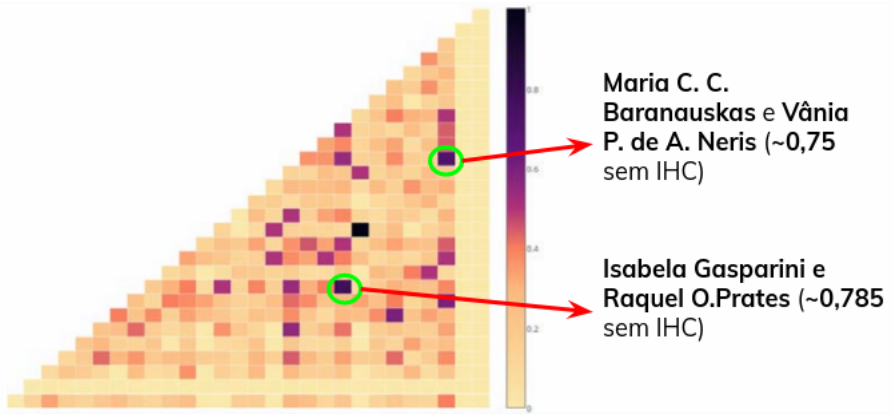
Figura 4.19 – Heatmap contendo as similaridades para cada par de autores



Fonte: Produção do autor.

Na Figura 4.20 podemos ver os pares mais similares encontrados que foram o par formado pelas autoras pela Maria C. C. Baranauskas e Vânia Paula de A. Neris (coeficiente aproximadamente igual a 0,75) e o par formado pelas autores Isabela Gasparini e Raquel O. Prates (coeficiente aproximadamente igual a 0,785).

Figura 4.20 – Heatmap contendo as similaridades para cada par de autores (mostrando os maiores coeficientes encontrados)



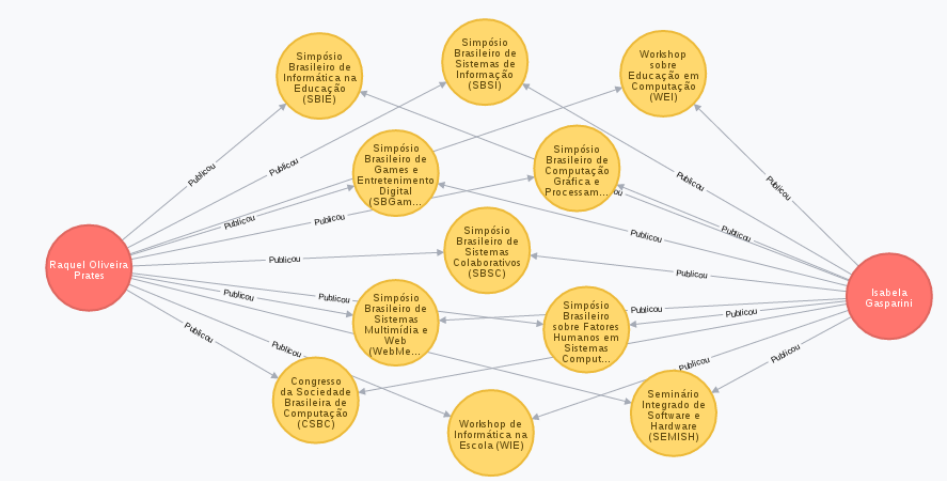
Fonte: Produção do autor.

Observando a Figura 4.20, pudemos notar que há um par que está mais escuro que todos, no limite, ou seja, valor 1 de similaridade, o que seria um par de autores que publicaram em todos veículos de publicação em comum, que seria o par formado pelos autores Roberto Pereira e Lara S. Godoy Piccolo (coeficiente igual a 1). Porém, ao verificar a lista de veículos da SBC que eles publicaram em comum, só havia um evento de publicação: o Simpósio Brasileiro de Informática na Educação (SBIE), ou seja, embora similares, foi somente em um evento. Já para o segundo maior valor de similaridade, no caso o par formado pela Isabela Gasparini e Raquel O. Prates foram encontrados 11 veículos de publicações diferentes conforme pode-se observar no grafo de similaridade na Figura 4.21, o que confere maior credibilidade a esta similaridade encontrada.

Coautores mais prolíficos

Observando apenas os coautores das publicações excluindo os 29 pesquisadores prolíficos, temos como verificar quais os autores mais prolíficos dentre estes coautores mostrados em um *"Top"10* na Tabela 4.11 sendo estes autores os que mais contribuíram em número de artigos com os 29 pesquisadores mais prolíficos de IHC.

Figura 4.21 – Grafo de Similaridade do par de autores Isabela Gasparini e Raquel O. Prates



Fonte: Produção do autor.

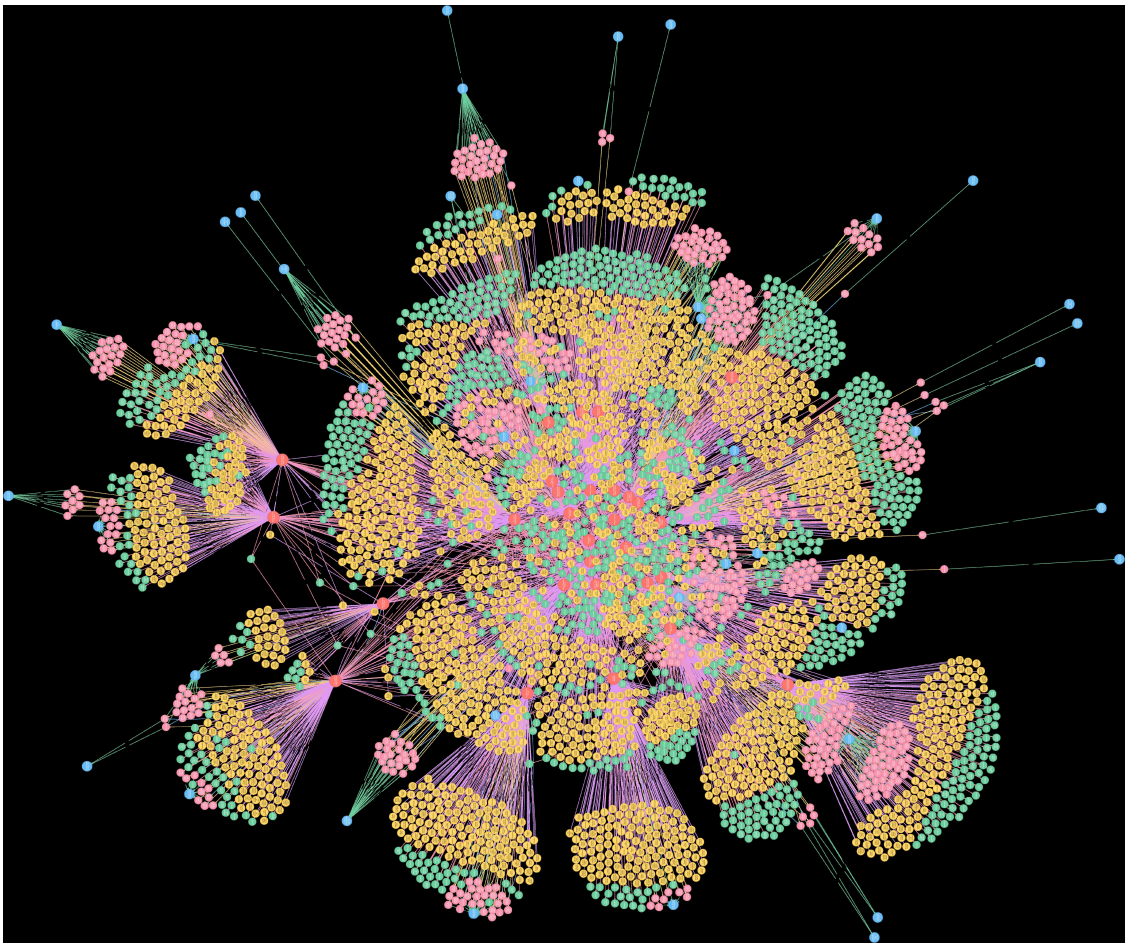
Tabela 4.11 – *Top* 10 Coautores mais prolíficos das publicações em parceria com os 29 autores mais prolíficos de IHC

Pos.	**Coautor**	**Qtde**
1	José Palazzo Moreira de Oliveira	61
2	Heiko Horst Hornung	58
	Philippe Palanque	
3	Aparecido Fabiano Pinatti de Carvalho	55
	Luciana Porcher Nedel	
4	Rodrigo Bonacin	52
5	Aline da Silva Alves	51
	Avanilde Kemczinski	
6	Elaine Cristina Saito Hayashi	43
7	Marcos Alexandre Rose Silva	42
8	Maria Cecília Martins	38
	Sidney Fels	
9	Ana Cristina Bicharra Garcia	35
10	Kátia Morosov Alonso	34

4.3.2 Análises de Redes Sociais

A Rede de Colaboração Científica da Comunidade Brasileira de IHC completa, como pode ser observada na Figura 4.22 foi construída a partir de 5 subredes: Relacionamentos de Coautoria entre os pesquisadores prolíficos e seus coautores; Relacionamento entre os pesquisadores prolíficos e seus orientados de mestrado ou doutorado, além das respectivas ligações com suas universidades tanto para os pesquisadores quanto para os orientados; Relacionamento entre os pesquisadores prolíficos e os veículos de publicação nos quais ele publicou.

Figura 4.22 – Rede de Colaboração Científica da Comunidade Brasileira de IHC completa



Fonte: Produção do autor.

Para ficar claro a distinção dos nós na rede, foram atribuídas cores para cada tipo de nó especificados na Tabela 4.12, e que serão usados em todas imagens desta seção.

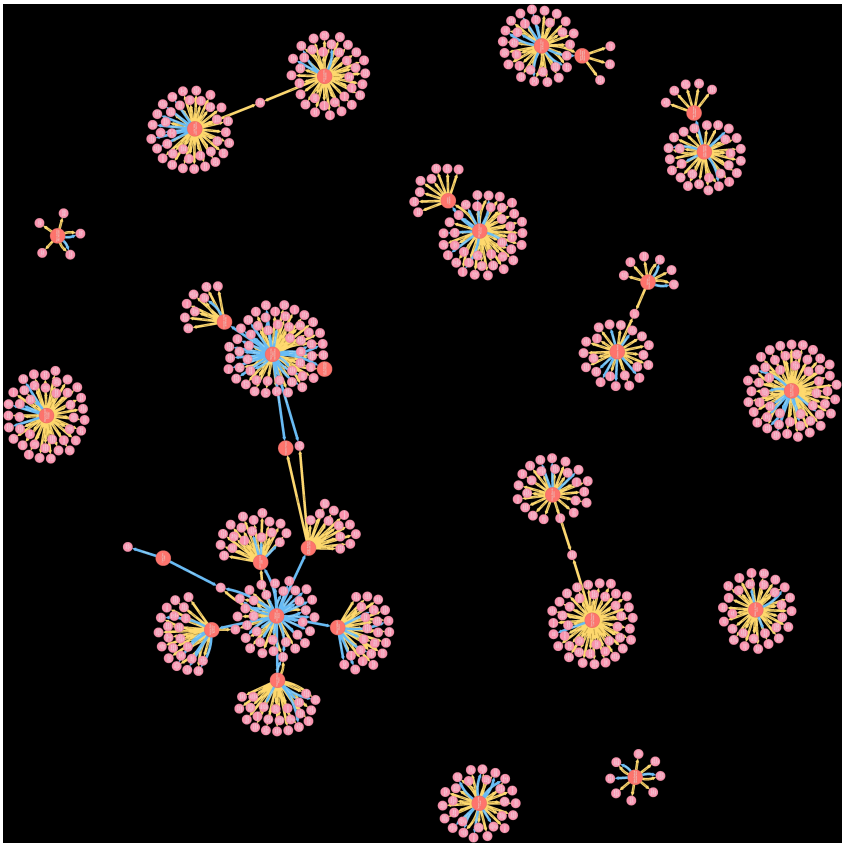
Tabela 4.12 – Legenda de cores nos grafos apresentados nesta seção

Cor	Tipo de Nó
Azul	Instituição/Universidade
Verde	Veículo de Publicação
Vermelho	Pesquisador Prolífico
Amarelo	Coautor
Rosa	Orientados de Mestrado e Doutorado

Visualização e Análises das Redes Sociais Geradas

A partir da Rede de Colaboração Científica mostrada na Figura 4.22, podemos visualizar as subredes formadas, como os relacionamentos entre os pesquisadores e seus orientados de mestrado e doutorado mostrado na Figura 4.23.

Figura 4.23 – Rede de Orientados de Mestrado e Doutorado dos autores mais prolíficos de IHC

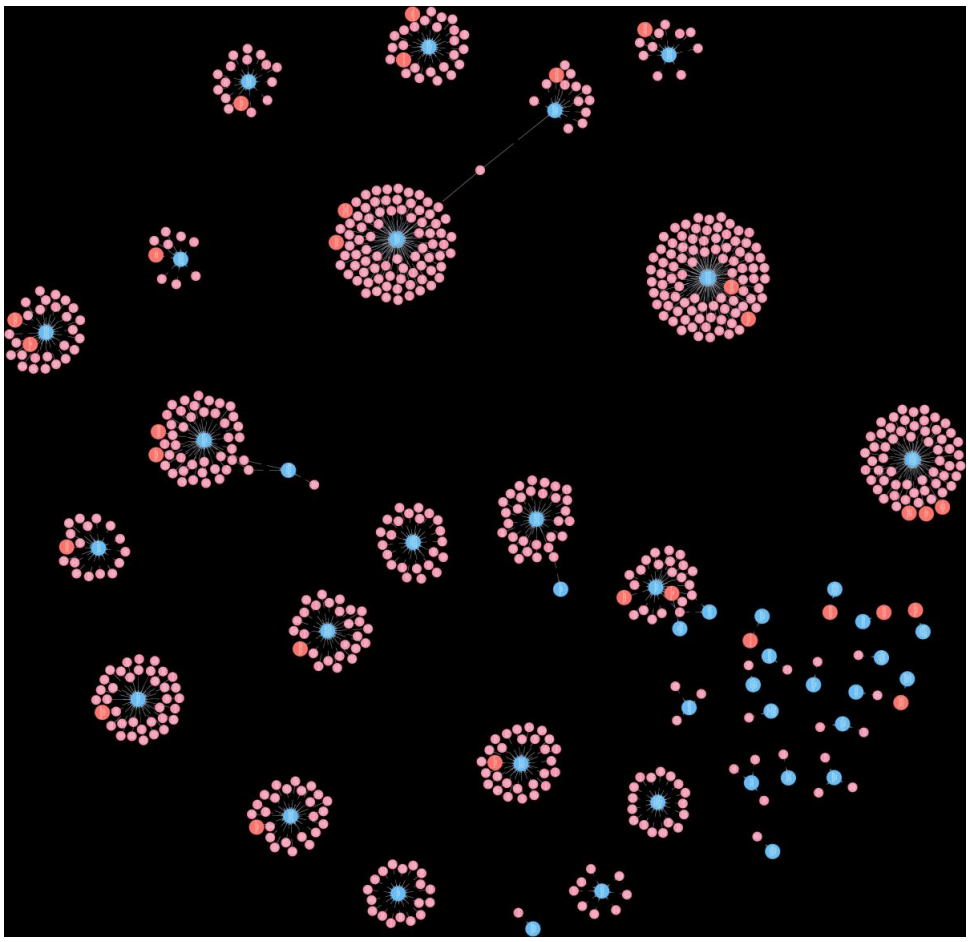


Fonte: Produção do autor.

É possível visualizar que existem relações de orientações de Mestrado e de Doutorado entre os próprios pesquisadores prolíficos (As relações de Mestrado foram coloridas de Amarelo, enquanto as relações de Doutorado foram coloridas com a cor Azul). Partindo disso, vemos que são formados diversos pares de pesquisadores e orientados, inclusive com alguns pesquisadores com orientados em comum, além de uma grande ilha

formada por 11 pesquisadores, com o nó central representado por uma das pesquisadores mais influentes: Clarisse Sieckenius de Souza que orientou diversos pesquisadores de Doutorado que se tornaram autores prolíficos dentro da comunidade de IHC.

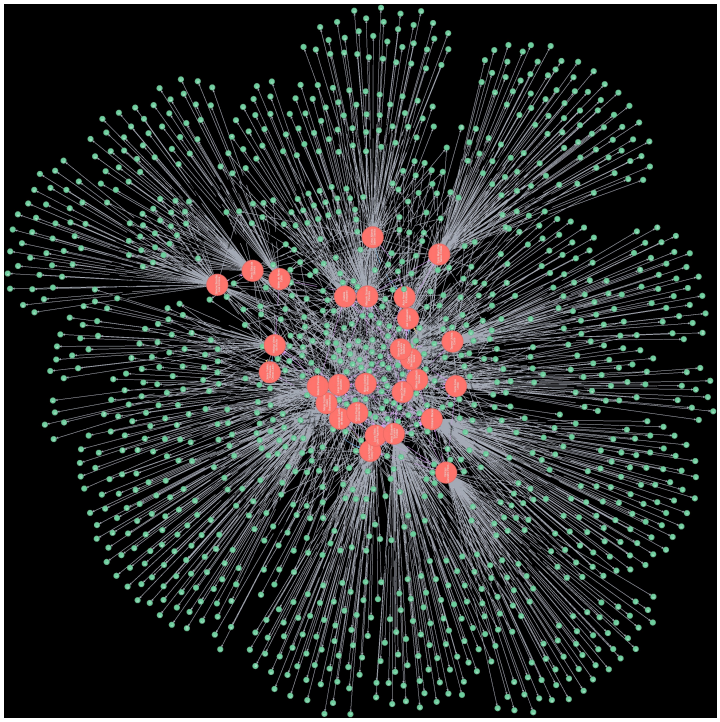
Figura 4.24 – Rede de Orientados e Pesquisadores e suas respectivas Universidades



Fonte: Produção do autor.

Na Figura 4.24 é possível ver que existem poucos orientados que fizeram o Mestrado e o Doutorado em Instituições diferentes, além do que existem poucas Universidades com mais de um Pesquisador Prolífico. Entre as Universidades que mais de destacam pela quantidade de orientados estão a Pontifícia Universidade Católica do Rio de Janeiro(PUC-Rio), a Universidade Federal do Rio Grande do Sul(UFRGS), e a Universidade Estadual de Campinas (UNICAMP). Já na Figura 4.25 é possível visualizar a grande quantidade de veículos de publicação que foram encontrados nas publicações dos 29 autores, ao total: 1199 Veículos de Publicação distintos.

Figura 4.25 – Rede de Pesquisadores e os veículos de publicação nos quais cada um publicou

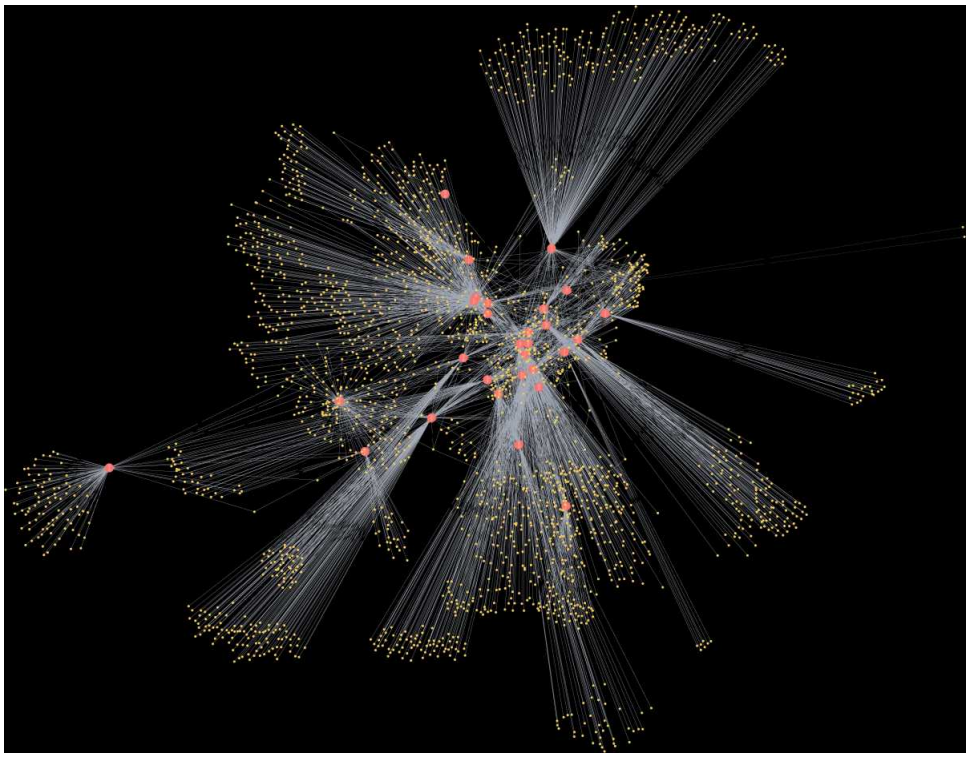


Fonte: Produção do autor.

Coautor com maior número de coautorias distintas

A principal subrede e também o grafo com maior número de elementos foi a Rede de Coautoria entre os pesquisadores e os autores que realizaram parceria em suas publicações mostrado na Figura 4.26. Dentro da rede é possível visualizar que existem muitos autores deslocados nos quais publicaram somente em parceria com um dos autores mais prolíficos, porém há um conjunto de autores na parte central do grafo que fizeram parceria com mais de um autor prolífico. Para isto foi realizado uma consulta no Neo4j para que pudesse ser visto quais seriam os autores que realizaram mais parcerias em número de conexões como pode ser visto na Tabela 4.13 o "*Top* 3"dos autores que mais realizaram coautorias distintas. Na Figura 4.27 é mostrado com detalhe o autor com mais conexões de coautoria.

Figura 4.26 – Rede de Coautoria dos Pesquisadores mais prolíficos de IHC

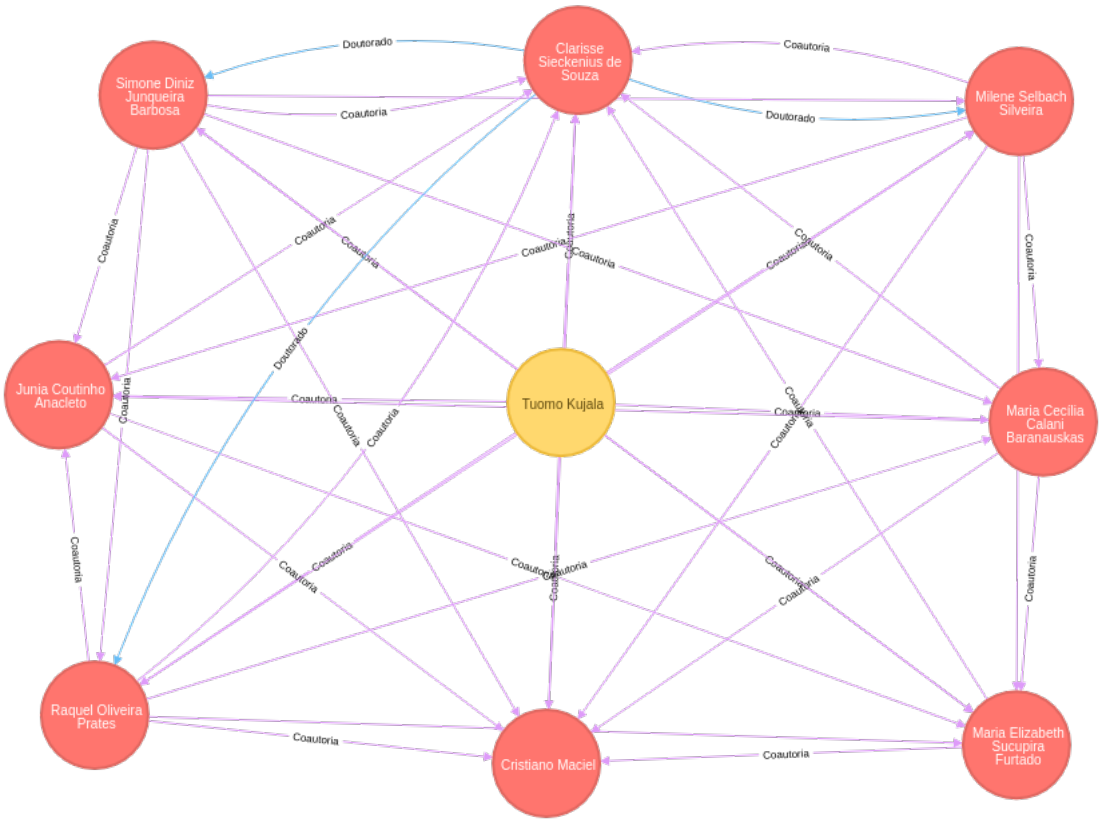


Fonte: Produção do autor.

Tabela 4.13 – Autores com maior número de coautorias distintas

#	**Autor**	**Nº de Coautorias Distintas**
1	Tuomo Kujala	8
	Paulo Melo	
	Heiko Horst Hornung	
	Clodis Boscarioli	
2	Elton José da Silva	7
3	Alberto Barbosa Raposo	6
	Adriana Holtz Betiol	
	Philippe Palanque	
	Maria Cecília Martins	

Figura 4.27 – Autor com maior número de coautorias distintas



Fonte: Produção do autor.

Os 8 pesquisadores prolíficos ligados ao autor com maior número de coautorias distintas Tuomo Kujala, não só tem essa ligação em comum, como são altamente interligados entre todos em forma de coautoria, além de existir algumas orientações de Doutorado e Mestrado na rede. É interessante ver a comparação entre a Tabela 4.11 e a Tabela 4.13, em que a primeira apresentou o autor mais prolífico dentre os coautores, com base no número de artigos, enquanto a segunda tabela apresentou o autor com maior número de coautorias distintas. Vemos que o coautor mais prolífico José Palazzo Moreira de Oliveira com 61 publicações em coautoria com os pesquisadores prolíficos nem aparece no *Top* 3 em número de coautorias distintas, já que ele publicou em parceria com apenas 3 pesquisadores, enquanto que os dois segundos lugares da primeira tabela, Heiko Horst Hornung e Philippe Palanque com 58 publicações cada um, aparecem na segunda tabela com 8 e 6 respectivamente, coautorias distintas.

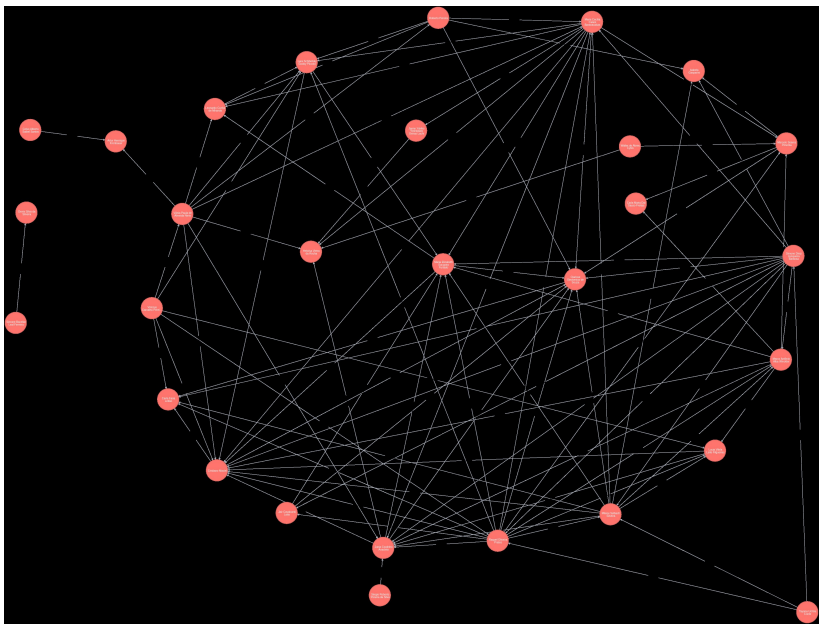
Ainda com base na Rede de Coautoria dos Pesquisadores Prolíficos, podemos ver a coautoria entre eles, retirando os autores que não sejam os mais prolíficos, sobrando apenas os 29 pesquisadores principais como pode ser visualizado na Figura 4.28. É possível visualizar que os 29 pesquisadores exercem uma colaboração científica muito forte entre

Tabela 4.14 – Métricas de Centralidade para a Rede de Coautoria dos Pesquisadores Prolíficos

Rede de Coautoria - Medidas de Centralidade					
Centralidade de Grau		**Centralidade de Proximidade**		**Centralidade de Intermediação**	
Autor	**Score**	**Autor**	**Score**	**Autor**	**Score**
Cristiano Maciel	228	Raquel Oliveira Prates	0.024	Raquel Oliveira Prates	4096
Raquel Oliveira Prates	186	Cristiano Maciel	0.023	Maria Cecília Calani Baranauskas	4163
Marcelo Soares Pimenta	180	Maria Cecília Calani Baranauskas	0.023	Cristiano Maciel	3825
Milene Selbach Silveira	177	Clarisse Sieckenius de Souza	0.022	Junia Coutinho Anacleto	3674
Isabela Gasparini	170	Milene Selbach Silveira	0.022	Milene Selbach Silveira	3531
Carla Maria Dal Sasso Freitas	167	Simone Diniz Junqueira Barbosa	0.022	Marcelo Soares Pimenta	3342
Junia Coutinho Anacleto	162	Junia Coutinho Anacleto	0.021	Simone Diniz Junqueira Barbosa	3306
Maria Cecília Calani Baranauskas	153	Maria Elizabeth Sucupira Furtado	0.020	Isabela Gasparini	2087
Tayana Uchôa Conte	148	Vânia Paula de Almeida Neris	0.020	Marco Antônio Alba Winckler	1996
Lucia Vilela Leite Filgueiras	144	Marcelo Soares Pimenta	0.019	Maria Elizabeth Sucupira Furtado	1960

eles, com algumas poucas exceções.

Figura 4.28 – Rede de Coautoria dos Pesquisadores mais prolíficos de IHC somente com os Pesquisadores Prolíficos



Fonte: Produção do autor.

Medidas de Centralidades baseadas nas Análises de Redes Sociais

Baseando nas métricas apresentadas no Capítulo 2, foram calculadas diversas métricas com base das redes geradas, entre elas: a Centralidade de Grau, Centralidade de Intermediação e Centralidade de Proximidade. Na Tabela 4.14 podemos ver essas métricas para a Rede de Coautoria dos Pesquisadores Prolíficos. Percebe-se que os autores que se destacaram nas três centralidades foram os autores Cristiano Maciel e Raquel Oliveira Prates, enquanto que outros autores como Maria C. C. Baranauskas, Marcelo Soares Pimenta divergem em algumas centralidades.

Tabela 4.15 – Métricas de Centralidade para a Rede de Orientações para Doutorado dos Pesquisadores Prolíficos

Rede de Orientações de Doutorado - Medidas de Centralidade					
Centralidade de Grau		**Centralidade de Proximidade**		**Centralidade de Intermediação**	
Autor	**Score**	**Autor**	**Score**	**Autor**	**Score**
Clarisse Sieckenius de Souza	29	Celso Alberto Saibel Santos	1	Clarisse Sieckenius de Souza	136
Maria Cecília Calani Baranauskas	24	Heloísa Vieira da Rocha		Maria Cecília Calani Baranauskas	35.5
Carla Maria Dal Sasso Freitas	11	Marcelo Soares Pimenta		Milene Selbach Silveira	18
Marcelo Soares Pimenta	10	Maria Cecília Calani Baranauskas	0.333	Simone Diniz Junqueira Barbosa	
Heloísa Vieira da Rocha	8	Leonardo Cunha de Miranda	0.2	Raquel Oliveira Prates	12
Tayana Uchôa Conte		Clarisse Sieckenius de Souza	0.142	Jair Cavalcanti Leite	9
Simone Diniz Junqueira Barbosa	7	Jair Cavalcanti Leite	0.833	Marcelo Soares Pimenta	4.5
Lucia Vilela Leite Filgueiras		Milene Selbach Silveira		Heloísa Vieira da Rocha	3.5
Milene Selbach Silveira		Raquel Oliveira Prates		Carla Faria Leitão	3
Raquel Oliveira Prates	6	Simone Diniz Junqueira Barbosa		Celso Alberto Saibel Santos	2

Tabela 4.16 – Métricas de Centralidade para a Rede de Orientações para Mestrado dos Pesquisadores Prolíficos

Rede de Orientações de Mestrado - Medidas de Centralidade					
Centralidade de Grau		**Centralidade de Proximidade**		**Centralidade de Intermediação**	
Autor	**Score**	**Autor**	**Score**	**Autor**	**Score**
Carla Maria Dal Sasso Freitas	44	Heloísa Vieira da Rocha	1	Clarisse Sieckenius de Souza	85
Simone Bacellar Leal Ferreira	40	Isabela Gasparini		Raquel Oliveira Prates	31.5
Maria Cecília Calani Baranauskas	40	Janne Yukiko Yoshikawa Oeiras Lachi		Simone Diniz Junqueira Barbosa	30
Maria Elizabeth Sucupira Furtado	36	Marcelo Soares Pimenta		Jair Cavalcanti Leite	22.5
Junia Coutinho Anacleto	33	Maria Cecília Calani Baranauskas		Maria Cecília Calani Baranauskas	19.5
Marcelo Soares Pimenta	32	Sérgio Roberto Pereira da Silva		Simone Bacellar Leal Ferreira	
Lucia Vilela Leite Filgueiras	30	Cristiano Maciel	0.5	Junia Coutinho Anacleto	16
Walter de Abreu Cybis	27	Denis da Silva Silveira		Marcelo Soares Pimenta	15.5
Heloísa Vieira da Rocha	26	Junia Coutinho Anacleto		Lucia Vilela Leite Filgueiras	14.5
Celso Alberto Saibel Santos	25	Lucia Vilela Leite Filgueiras		Heloísa Vieira da Rocha	12.5

Na Tabela 4.15 são apresentadas as métricas de centralidade para a Rede de Orientações para Doutorado e na Tabela 4.16 para a Rede de Orientações para Mestrado. Nota-se que aqui, os autores mais influentes seguindo a Rede de Orientações de Mestrado e Doutorado foram as pesquisadoras Maria C. C. Baranauskas e a Clarisse Sieckenius de Souza, duas pioneiras da área de IHC no Brasil.

Além de medir as centralidades dos autores mais prolíficos para descobrir quais seriam os mais influentes segundo as suas produções como orientações de mestrado e doutorado, e também com suas coautorias, foram medidas as centralidades de grau e de intermediação para as Universidades. Na Tabela 4.17 são apresentadas as Universidades que se mostraram centrais dentro da Rede de Colaboração Científica conforme as medidas de centralidade.

Por fim, foi medida a centralidade dos autores considerando toda a Rede de Colaboração Científica, ou seja, as centralidades dos autores considerando suas orientações, coautorias, universidades e veículos de publicações, como pode ser visto na Tabela 4.18. Levando em conta os resultados, existem muitos autores influentes, mas o nome de destaque nas centralidades foi a autora Maria Cecília Calani Baranauskas.

Tabela 4.17 – Métricas de Centralidade para as Universidades da Rede de Colaboração Científica

Rede de Vínculos as Instituições - Medidas de Centralidade			
Centralidade de Grau		**Centralidade de Intermediação**	
Instituição	**Score**	**Instituição**	**Score**
Universidade Federal do Rio Grande do Sul	90	Universidade Estadual de Campinas	40
Universidade Estadual de Campinas	81	Universidade Federal de São Carlos	20.5
Pontíficia Universidade Católica do Rio de Janeiro	59	Universidade Federal de Santa Catarina	14.5
Universidade Federal de São Carlos	43	Universidade Salvador	10
Universidade de Fortaleza	36	Universidade Estadual de Maringá	6.5

Tabela 4.18 – Métricas de Centralidade para a Rede de Colaboração Científica Completa dos Pesquisadores Prolíficos

Rede de Colaboração Científica Completa - Medidas de Centralidade					
Grau		**Proximidade**		**Intermediação**	
Autor	**Score**	**Autor**	**Score**	**Autor**	**Score**
Maria Cecília Calani Baranauskas	398	Raquel Oliveira Prates	0.0243	Cristiano Maciel	4847
Cristiano Maciel	389	Maria Cecília Calani Baranauskas	0.0238	Maria Cecília Calani Baranauskas	4835
Marcelo Soares Pimenta	361	Cristiano Maciel	0.0232	Junia Coutinho Anacleto	4197
Carla Maria Dal Sasso Freitas	322	Clarisse Sieckenius de Souza	0.0227	Marcelo Soares Pimenta	3948
Junia Coutinho Anacleto	312	Junia Coutinho Anacleto	0.0227	Raquel Oliveira Prates	3597
Raquel Oliveira Prates	293	Milene Selbach Silveira	0.0227	Carla Maria Dal Sasso Freitas	3577
Maria Elizabeth Sucupira Furtado	288	Simone Diniz Junqueira Barbosa	0.0227	Maria Elizabeth Sucupira Furtado	3466
Lucia Vilela Leite Filgueiras	282	Maria Elizabeth Sucupira Furtado	0.0222	Milene Selbach Silveira	3357
Milene Selbach Silveira	280	Marco Antônio Alba Winckler	0.0212	Lucia Vilela Leite Filgueiras	3249
Simone Diniz Junqueira Barbosa	272	Vânia Paula de Almeida Neris	0.0212	Clarisse Sieckenius de Souza	3041

PageRank

Além das medidas de centralidade mais tradicionais de grafos apresentadas, temos uma medida de centralidade variante da centralidade de Auto Vetor ou Vetor Próprio que é o PageRank. O PageRank avalia um membro da rede baseando na quantidade, qualidade e contexto das ligações que uma pessoa recebe e faz (FáBIORICOTTA, 2016). Nas Tabelas 4.19 e 4.20 foram calculados os valores dos nós do tipo “Pesquisador”, “Veículos de Publicação”e “Instituições”usando a métrica do PageRank.

Tabela 4.19 – Medida de Page Rank para os nós Pesquisador e Universidade

Medida de PageRank para os nós Pesquisador e Universidade			
Pesquisador	**Score**	**Universidade**	**Score**
Cristiano Maciel	25.93	Universidade Federal do Rio Grande do Sul	19.14
Milene Selbach Silveira	19.21	Universidade Estadual de Campinas	14.01
Isabela Gasparini	19.16	Pontifícia Universidade Católica do Rio de Janeiro	11.09
Carla Maria Dal Sasso Freitas	18.91	Universidade Federal de São Carlos	8.89
Raquel Oliveira Prates	17.90	Universidade Federal do Amazonas	7.95
Tayana Uchôa Conte	17.62	Universidade de Fortaleza	6.63
Marcelo Soares Pimenta	17.42	Pontifícia Universidade Católica do Rio Grande do Sul	6.21
Lucia Vilela Leite Filgueiras	16.95	Universidade Federal de Minas Gerais	6.11
Junia Coutinho Anacleto	16.91	Universidade Federal do Rio Grande do Norte	6.04
Maria Elizabeth Sucupira Furtado	13.34	Universidade Federal de Santa Catarina	5.92

Tabela 4.20 – Medida de PageRank para o nó Veículo de Publicação

Medida de PageRank para o nó Veículo de Publicação	
Veículo de Publicação	**Score**
Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais (IHC)	2.77
International Conference on Human-Computer Interaction (HCII)	2.35
Congresso Latino-Americano de Interação Humano-Computador (CLIHC)	2.14
IFIP TC International Conference on Human-Computer Interaction (INTERACT)	2.02
Congresso da Sociedade Brasileira de Computação (CSBC)	1.94
Simpósio Brasileiro de Informática na Educação (SBIE)	1.90
Simpósio Brasileiro de Multimídia e Web (WebMedia)	1.71
Conferencia Latinoamericana en Informática (CLEI)	1.55
IADIS International Conference WWW/Internet (ICWI)	1.48
Journal of the Brazilian Computer Society	1.34

5 Conclusões

As redes sociais vêm adquirindo extrema importância no mundo atual, onde com o advento das mídias sociais trazido pela internet tem conseguido criar redes sociais nunca antes imaginadas, como por exemplo, as redes sociais formadas no meio acadêmico. As técnicas de análise de redes sociais permitem que estas conexões sejam percebidas e possibilita tirar conclusões relevantes sobre estas ligações.

Para elaboração deste trabalho, houve a necessidade de conhecer os conceitos sobre diversas técnicas de análises de informações como análises bibliométricas que ajudam a construir métricas para diversos dados e retirar informações importantes. Para que seja gerada a rede de colaboração científica, também é necessário saber sobre o material de estudo, no caso, a Comunidade Brasileira de IHC, formada por estudantes, pesquisadores e professores da área de Interação Humano-Computador, as quais são as pesquisas já realizadas e como a comunidade tem se comportado até então. Além disso, é necessário saber os fundamentos das redes sociais e suas vertentes como a rede de colaboração científica que é o tema deste trabalho, e como consequência, a análise de redes sociais é necessária para que a rede possa ser analisada utilizando diversas métricas apresentadas. Por fim, foram mostrados os conceitos por trás do processo de Extração de Conhecimento (KDD), e como ele pode ser utilizado para o entendimento, refinamento, e processamento dos dados que irão ser trabalhados usando diversas técnicas da área de extração de conhecimento.

Nos trabalhos relacionados, foram apresentados os trabalhos que já realizaram pesquisas sobre análise de redes sociais ligadas de alguma forma ao mundo acadêmico, e também na extração de informações e processamento dos dados dos currículos Lattes. Entre eles, diversos estudos relevantes foram feitos, com diversos fins, mas que irão auxiliar no embasamento e guiar este trabalho para o seu objetivo que é gerar as redes de colaboração científica da comunidade brasileira de IHC.

Por fim, foi apresentada o trabalho realizado deste Trabalho de Conclusão de Curso, e quais etapas foram feitas desde a extração dos dados dos currículos da Plataforma Lattes e a sua inserção no banco de dados para divisão em tabelas. Também foi realizada a implementação da ferramenta em PHP que realizou a extração automática

dos dados dos currículos em formato XML para serem armazenados neste banco de dados. Assim foi realizado a filtragem, padronização e limpeza das informações obtidas, de modo que os dados ficassem adequados para realizar as análises bibliométricas. Além do uso das tabelas nas análises bibliométricas, as tabelas puderam ser utilizadas para inserir os dados e gerar a Rede de Colaboração Científica da Comunidade Brasileira de IHC, e por fim, analisar com métricas de redes sociais. É importante salientar que boa parte das análises realizadas nesta pesquisa culminaram na publicação de um artigo no próprio Simpósio tema desta pesquisa, o XVI Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais, com o título: "*Crossing the borders of IHC: where else have our researchers been publishing?*"onde foi mostrado diversas análises feitas neste trabalho de conclusão de curso.

Por meio das análises realizadas, foi possível verificar quais os principais fóruns os autores da comunidade têm publicado (dentro e fora do IHC), além de verificar uma maior presença das publicações em veículos internacionais. Foi analisado os tipos de publicações, similaridades entre os autores, entre outras reflexões que servem para ajudar entender e moldar o perfil da comunidade brasileira de IHC. Também foi comprovado a interdisciplinaridade já que foram encontrados veículos de publicação de diversos fóruns, não só da área de IHC. Aliás, houve a constatação de uma escassez de periódicos brasileiros que tratem especificamente da área de Interação Humano-Computador, porém, apesar disso, há um esforço contínuo da comunidade de se tornar visível internacionalmente. É esperado que a reflexão sobre a comunidade traga maior qualidade para as discussões e ajude a traçar estratégias para aumentar a inserção internacional.

Além das análises, tivemos a própria geração da Rede de Colaboração Científica da Comunidade Brasileira de IHC, a análise da proximidade entre os autores e a produtividade e influência de cada um dentro da rede formada. Pudemos constatar também os coautores mais prolíficos dentro da rede de coautoria dos autores mais prolíficos de IHC, tanto com base no número de artigos em coautoria, quanto no número de coautorias distintas.

Entre as limitações do trabalho realizado foram a falta de padronização de nomes de pesquisadores, instituições e veículos de publicações que poderia ser resolvido ou amenizado caso a Plataforma Lattes tivesse um banco de dados prévio pelo menos das principais instituições e veículos de publicações, pois, dessa forma, quando um usuário fosse cadastrar um novo artigo por exemplo, ele já poderia escolher em um seletor a

universidade e o veículo de publicação escolhido com os nomes padronizados e corretos. Outra limitação é alguns campos vazios em algumas informações das publicações o que deixa algumas análises incompletas. Essa fase de limpeza e padronização também impediu que fosse realizado uma atualização dos dados no decorrer da pesquisa, pois, mesmo que fosse realizado uma nova captura dos dados atualizados do Lattes, e das publicações do IHC 2016, teria que refazer toda a padronização e limpeza das informações, que foi justamente o maior tempo gasto no andamento do trabalho.

5.1 Trabalhos Futuros

Como trabalhos futuros, é possível expandir a busca nos CV Lattes para todos os pesquisadores da área que já publicaram no IHC. Além disso, também poderia ser complementada com a busca pelos pesquisadores nos anais das principais conferências identificadas na pesquisa. Outra análise interessante seria adaptar o Número de Erdős apresentado no Capítulo 2 para o(s) autor(es) mais influentes de IHC encontrados nos resultados apresentados no Capítulo 4, podendo assim medir a distância de um pesquisador para o autor mais influente de IHC com base nas coautorias das publicações.

Na rede formada também poderia ter sido realizada a detecção de grupos ou comunidades. Esta análise tem como objetivo definir grupos de pesquisadores que possuem mais conexões entre si do que com que com o restante da rede. A detecção das comunidades pode ser obtida pelo próprio banco de dados Neo4j, já que ele conta com alguns *plugins* ou bibliotecas extras que podem ser instaladas, e assim, utilizar algoritmos de pesquisas e análises como a detecção de comunidades e centralidades por exemplo. Porém, para a realização desta análise na rede de colaboração científica necessitaria da especificação de "pesos"em relação aos nós, e também, em relação aos relacionamentos, o que teria que ser padronizado de acordo com algum padrão previamente criado, para então, com a rede valorada, ser detectado as comunidades. Nas redes formadas ainda poderiam ser caracterizadas quanto à sua região geográfica caso fossem extraídas estas informações, realizar uma análise da coautoria ao longo do tempo, e também uma análise mais profunda em relação a coautoria internacional.

Em relação as limitações encontradas nesta pesquisa, poderia ser realizado a criação de um *script* ou melhoria do *script* já implementado neste trabalho para realizar

a desambiguação de nomes encontrados nos currículos, dessa forma, tornando o trabalho mais automático, e consequentemente, mais rápido.

Referências

ALMEIDA, A. *Trabalhando com Relacionamentos: bancos de dados baseados em grafos e o Neo4j*. 2011. `http://blog.caelum.com.br/trabalhando-com-relacionamentos-bancos-de-dados-baseados-em-\grafos-e-o-neo4j`. Accessed: 2017-10-30.

ALVARADO, R. U. A bibliometria no brasil. *Ciência da Informação*, v. 13, n. 2, p. 91–105, 1984.

ALVES, A. D.; YANASSE, H. H.; SOMA, N. Y. Sucupira: a system for information extraction of the lattes platform to identify academic social networks. In: *Iberian Conference on Information Systems and Technologies (CISTI)*. Chaves, Portugal: , 2011.

AMORIM, T. *Conceitos, técnicas, ferramentas e aplicações de Mineração de Dados para gerar conhecimento a partir de bases de dados*. 2006. Monografia (Bacharel em Ciência da Computação), UFPE (Universidade Federal de Pernambuco), Pernambuco, Brasil.

ARAO, L. H. *Vida média e Obsolência da área de Ciência da Literatura: uma contribuição ao entendimento da cronologia de citações na atividade acadêmica*. Rio de Janeiro: UFRJ, 2014. Trabalho de Conclusão de Graduação.

ARAUJO, C. A. Bibliometria: evolução histórica e questões atuais. *Em Questão*, v. 12, n. 1, p. 11–32, 2006.

BALANCIERI, R. *Análise de Redes de Pesquisa em uma Plataforma de Gestão em Ciência e Tecnologia: uma aplicação à Plataforma Lattes*. Florianópolis: UFSC, 2004. Dissertação de Mestrado.

BARABáSI, A. L.; JEONG, H.; NéDA, Z.; RAVASZ, E.; SCHUBERT, A.; VICSEK, T. Evolution of the social network of scientific collaborations. *Journal Physica A: Statistical Mechanics and its Applications*, p. 590–614, 2001.

BARBOSA, S. D. J.; SILVA, B. S. da. *Interação Humano-Computador*. Rio de Janeiro, Brasil: Elsevier, 2010.

BARBOSA, S. D. J.; SILVEIRA, M. S.; GASPARINI, I. What publications metadata tell us about the evolution of a scientific community: the case of the brazilian human–computer interaction conference series. *Scientometrics*, v. 110, p. 275–300, 2017.

BENEVENUTO, F.; ALMEIDA, J.; SILVA, A. S. da. Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações. In: *Minicurso in XXIX Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. Campo Grande, Mato Grosso do Sul: , 2011.

BUENO, M. F.; VIANA, M. R. Mineração de dados: Aplicações, eficiência e usabilidade. *Congresso de Iniciação Científica do INATEL*, 2012.

BURKE, F. *Social Media vs. Social Networking*. 2013. http://www.huffingtonpost.com/fauzia-burke/social-media-vs-social-ne_b_4017305.html. Accessed: 2016-10-26.

CARLINI-GARCIA, L. A. *Estudo da estrutura genética populacional através de marcadores moleculares*. Piracicaba, Escola Superior de Agricultura "Luiz de Queiroz": USP, 1998. Monografia (Pós-Graduação).

CARVALHO, L. A. V. de. Data mining – a mineração de dados no marketing, medicina, economia, engenharia e administração. *Ciência Moderna*, 2005.

CHATTI, M. A.; DYCHKHOFF, A. L.; SCHROEDER, U.; THUS, H. A reference model for learning analytics. In: *International Journal of Technology Enhanced Learning*. Paris, França: , 2012. v. 4, n. 5/6, p. 318–331.

CIRIBELI, J. P.; PAIVA, V. H. P. Redes e mídias sociais na internet: realidades e perspectivas de um mundo conectado. *Revista Mediação*, v. 13, n. 12, 2011.

DANUELLO, J. C. *Produção científica docente em tratamento temático da informação no Brasil: uma abordagem métrica como subsídio para a análise do domínio*. 2007. (Dissertação de Mestrado) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista.

DIGIAMPETRI, L. A.; MENA-CHALCO, J. P.; PEREZ-ALCáZAR, J. J.; TUESTA, E. F.; DELGADO, K. V.; MUGNAINI, R.; SILVA, G. S. Minerando e caracterizando dados de currículos lattes. In: *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. Curitiba, Brasil: , 2012.

DIGIAMPETRI, L. A.; SILVA, E. E. da. A framework for a social network of researchers analysis. In: *Iberoamerican Journal of Applied Computing*. 2011.

EGGHE, L.; ROUSSEAU, R. *Introduction to informetrics: quantitative methods in library, documentation and information science*. Amsterdam: Elsevier Science, 1990.

FERREIRA, E. R.; JúNIOR, S. M. T. *Analise de desempenho de Bancos de Dados*. 2012. Monografia (Bacharel em Ciências da Computação), UNIPAC (Universidade Presidente Antônio Carlos), Barbacena, Brasil.

FORTUNATO, S. Community detection in graphs. In: *Physics report*. 2010.

FOUNDATION, P. S. *Python 3.6.3 documentation*. 2017. https://docs.python.org/3/faq/general.html#what-is-python. Accessed: 2017-10-30.

FREITAS, L. Q. de. *Medidas de centralidade de grafos*. Rio de Janeiro: UFRJ, 2010. Tese de Doutorado.

FáBIORICOTTA. *O que é PageRank?* 2016. http://www.agenciamestre.com/marketing-digital/o-que-e-pagerank/. Accessed: 2016-10-30.

GABARDO, A. C. *Análise de redes sociais: uma visão computacional*. 1a. ed. São Paulo: Novatec Editora, 2015.

GASPARINI, I.; BARBOSA, S. D. J.; SILVEIRA, M. S.; BIM, S. A.; BOSCARIOLI, C. How does hci research affect education programs? a study in the brazilian context. In: *Proceedings of the Human-Computer Interaction – INTERACT 2015*. Bamberg, Alemanha: , 2015.

GASPARINI, I.; CUNHA, L. F. da; KIMURA, M. H.; PIMENTA, M. S. Análise das redes de coautoria do simpósio brasileiro sobre fatores humanos em sistemas computacionais. In: *Proceedings of the 13th Brazilian Symposium on Human Factors in Computing Systems*. Foz do Iguaçu, Brasil: , 2014. p. 323–332.

GOLDBECK, J. *Introduction to Social Media Investigation: A Hands-On Approach*. Massachusetts: Elsevier, 2015.

GUEDES, V. L. S.; BORSCHIVER, S. Bibliometria: Uma ferramenta estatística para a gestão de informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. *Encontro Nacional de Ciência da Informação*, v. 6, p. 1–18, 2005.

HIRSCH, J. E. Does the h index have predictive power? In: *Proceedings of the National Academy of Sciences*. 2007.

JONES, M. *Introduction to HCI*. 2016. `https://www.cs.bham.ac.uk/~rxb/Teaching/HCI%20II/intro.html`. Accessed: 2016-10-30.

JúNIOR, A. G. da S. *Análise de Redes Sociais aplicada à Assistência Técnica e Extensão Rural (ATER)*. 2016. `http://pt.slideshare.net/equipeagroplus/agroplus-social-networkanalysis`. Accessed: 2016-10-30.

KATZ, S. J.; MARTIN, B. R. *What is research collaboration?* New York, EUA: Research policy, 1997. 1-18 p.

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

MARUYAMA, W. T.; DIGIAMPETRI, L. Predição de relacionamentos em redes sociais, uma revisão sistemática. In: *V Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2016)*. Porto Alegre, Brasil: , 2016.

MCGRATH, W. What bibliometricians, scientometricians and informetricians study; a typology for definition and classification; topics for discussion. In: *International Conference on Bibliometrics, Scientometrics and Informetrics*. Ontario, Canada: , 1989.

MENA-CHALCO, J. P.; JUNIOR, C.; MARCONDES, R. Scriptlattes: an open-source knowledge extraction system from the lattes platform. In: *Journal of the Brazilian Computer Society*. Porto Alegre, Brasil: , 2009.

MYSQL. *10 Principais Motivos para Usar o MySQL como um Banco de Dados Incorporado*.

MYSQL. 2016. `https://www.mysql.com`. Accessed: 2016-10-30.

NARIN, F.; WHITLOW, E. S. *Measurement of scientific cooperation and coauthorship in CEC-Related areas of science*. : Office for Official Publications of the European Communities, 1990.

NEO4J. *The Internet-Scale Graph Platform*. 2017. `https://neo4j.com/product/?ref=hro`. Accessed: 2017-10-30.

NEWMAN, M. E. J. The structure of scientific collaboration networks. In: *Proceedings of the National Academy of Sciences*. 2001. v. 98, n. 2, p. 404–409.

NEWMAN, M. E. J. *Who is the best connected scientist a study of scientific coauthorship networks*. Berlin: Springer Berlin Heidelberg, 2004.

OLIVEIRA, A. A. de. *Análise de classificação - Roteiro em R*. 2016. http://ecovirtual.ib.usp.br/doku.php?id=ecovirt:roteiro:comuni:comuni_classr. Accessed: 2017-10-30.

OLIVEIRA, E. F. T. de; SANTAREM, L. G. da S.; SEGUNDO, J. E. S. Análise das redes de colaboração científica através do estudo das co-autorias, nos cursos de pós-graduação do brasil no tema tratamento temático da informação. In: *Actas del IX Congreso ISKO-España: nuevas perspectivas para la difusión y organización del conocimiento. Valencia: Sociedad Internacional Para La Organización del Conocimiento–Capítulo Español*. 2009. p. 309–327.

ORACLE. *Oracle MySQL*. 2016. https://www.oracle.com/br/mysql/index.html. Accessed: 2016-10-30.

PHP. 2016. http://php.net/. Accessed: 2016-10-30.

PHPMANUAL. 2016. https://secure.php.net/manual/pt_BR/index.php. Accessed: 2016-10-30.

PRICE, D. de S. *The structures of publication in science and technology*. Cambridge: MIT Press, 1969. 91-104 p.

PRITCHARD, A. Statistical bibliography or bibliometrics. *Journal of Documentation*, v. 25, n. 4, p. 348–349, 1969.

RIBEIRO, R. T. *SimpleXML para manipular XML pelo PHP*. 2016. http://rubsphp.blogspot.com.br/2011/02/simplexml.html?showComment=1380891173655. Accessed: 2016-10-30.

ROBINSON, I.; WEBBER, J.; EIFREM, E. *Graph Databases: New Opportunities for Connected Data*. : O'Reilly Media, 2015.

SANTOS, R. N. M. dos; KOBASHI, N. Y. Bibliometria, cientometria, infometria: Conceitos e aplicações. *Ciência da Informação*, v. 2, n. 1, p. 155–172, 2009.

SBC. *Grandes Desafios de Pesquisa em Interação Humano-Computador no Brasil. Relatório Técnico*. 2012. http://comissoes.sbc.org.br/ceihc/documentos/RT_GranDIHC_BR_2012.pdf. Accessed: 2016-10-30.

SCHOLAR, G. *Google Scholar Metrics*. 2017. https://scholar.google.com/intl/en/scholar/metrics.html. Accessed: 2917-10-30.

SCOTT, J. *Social network analysis: a handbook*. Exeter: Sage Publications, 2000.

SILVA, A. K. A. da; BARBOSA, R. R.; DUARTE, E. N. Rede social de coautoria em ciência da informação: estudo sobre a área temática de organização e representação do conhecimento". *Informação & Sociedade*, v. 22, n. 2, 2012.

SOUZA, C. S. de. *Da importância dos Simpósios Brasileiros de Fatores Humanos em Sistemas Computacionais*. 2006. http://comissoes.sbc.org.br/ce-ihc/documentos/da-importancia-dos-IHCs_2006.html. Accessed: 2016-10-30.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao Data Mining Mineração de Dados*. Rio de Janeiro: Ciência Moderna, 2009.

UDDIN, S.; HOSSAIN, L.; ABBASI, A.; RASMUSSEN, K. *Trend and efficiency analysis of co-authorship network*. : Scientometrics, 2012. 687-699 p.

VENNERS, B. *The Making of Python*. 2003. `http://www.artima.com/intv/pythonP.html`. Accessed: 2017-10-30.

VISWANATH, B.; MISLOVE, A.; CHA, M.; GUMMADI, K. P. On the evolution of user interaction in facebook. In: *Proceedings of the 2nd ACM workshop on Online social networks*. 2009.

W3C. *Extensible Markup Language (XML)*. 2016. `https://www.w3.org/XML/`. Accessed: 2016-10-30.

W3C. *XML Tutorial*. 2016. `http://www.w3schools.com/xml/default.asp`. Accessed: 2016-10-30.

WEISZ, J.; ROCO, M. C. Redes de pesquisa e educação em engenharia nas américas. In: *Rio de Janeiro: FINEP*. 1996.