UNIVERSIDADE DE PASSO FUNDO

INSTITUTO DE CIÊNCIAS EXATAS E GEOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

UMA ABORDAGEM PARA IDENTIFICAR A SIMILARIDADE ENTRE PERFIS DE PESQUISADORES COM VISTAS A RECOMENDAÇÃO

Diogo Nelson Rovadosky

Passo Fundo

2018

UNIVERSIDADE DE PASSO FUNDO
INSTITUTO DE CIÊNCIAS EXATAS E GEOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

UMA ABORDAGEM PARA IDENTIFICAR A SIMILARIDADE ENTRE PERFIS DE PESQUISADORES COM VISTAS A RECOMENDAÇÃO

Diogo Nelson Rovadosky

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Computação Aplicada na Universidade de Passo Fundo.

Orientador: Cristiano Roberto Cervi

Passo Fundo
2018

CIP – Catalogação na Publicação

R875a Rovadosky, Diogo Nelson
Uma abordagem para identificar a similaridade entre perfis de pesquisadores com vistas a recomendação / Diogo Nelson Rovadosky. – 2018.
124 f. : il. color. ; 30 cm.

Orientador: Prof. Dr. Cristiano Roberto Cervi.
Dissertação (Mestrado em Computação Aplicada) – Universidade de Passo Fundo, 2018.

1. Software. 2. Pesquisadores – Perfil. 3. Sistemas de recomendação. I. Cervi, Cristiano Roberto, orientador. II. Título.

CDU: 004.41

Catalogação: Bibliotecária Marciéli de Oliveira - CRB 10/2113

UPF Universidade de Passo Fundo | PPGCA Programa de Pós-Graduação em Computação Aplicada
Instituto de Ciências Exatas e Geociências - ICEG

ATA DE DEFESA DO TRABALHO DE CONCLUSÃO DE CURSO DO ACADÊMICO

DIOGO NELSON ROVADOSKY

Aos vinte e dois dias do mês de março do ano de dois mil e dezoito, às 9 horas, realizou-se, no Instituto de Ciências Exatas e Geociências, prédio B5, da Universidade de Passo Fundo, a sessão pública de defesa do Trabalho de Conclusão de Curso **"Uma abordagem para identificar similaridade entre perfis de pesquisadores com vistas a recomendação"**, de autoria de Diogo Nelson Rovadosky, acadêmico do Curso de Mestrado em Computação Aplicada do Programa de Pós-Graduação em Computação Aplicada – PPGCA/UPF. Segundo as informações prestadas pelo Conselho de Pós-Graduação e constantes nos arquivos da Secretaria do PPGCA, o aluno preencheu os requisitos necessários para submeter seu trabalho à avaliação. A banca examinadora foi composta pelos doutores Cristiano Roberto Cervi, Roberto dos Santos Rabello e Eduardo Kessler Piveta. Concluídos os trabalhos de apresentação e arguição, a banca examinadora considerou o candidato APROVADO. Foi concedido o prazo de até quarenta e cinco (45) dias, conforme Regimento do PPGCA, para o acadêmico apresentar ao Conselho de Pós-Graduação o trabalho em sua redação definitiva, a fim de que sejam feitos os encaminhamentos necessários à emissão do Diploma de Mestre em Computação Aplicada. Para constar, foi lavrada a presente ata, que vai assinada pelos membros da banca examinadora e pela Coordenação do PPGCA.

Prof. Dr. Cristiano Roberto Cervi - UPF
Presidente da Banca Examinadora
(Orientador)

Prof. Dr. Roberto dos Santos Rabello - UPF
(Avaliador Interno)

Prof. Dr. Eduardo Kessler Piveta - UFSM
(Avaliador Externo)

Prof. Dr. Rafael Rieder
Coordenador do PPGCA

AGRADECIMENTOS

Agradeço ao Prof. Dr. Cristiano Roberto Cervi, orientador, pela compreensão, confiança, paciência e valioso auxílio durante as etapas deste trabalho.

Agradeço também, a toda minha família que sempre apoiaram, incentivaram, com carinho e compreensão pelos momentos que deixei de participar em suas vidas.

Aos professores do PPGCA e aos colegas de turma, obrigado pelas discussões, auxílios e grande aprendizado durante o andamento das aulas.

E, por fim, agradeço a todos aqueles que de uma maneira ou outra colaboraram na realização deste trabalho.

UMA ABORDAGEM PARA IDENTIFICAR A SIMILARIDADE ENTRE PERFIS DE PESQUISADORES COM VISTAS A RECOMENDAÇÃO

RESUMO

A produção de conhecimento pela humanidade aumenta a cada dia que passa, gerando um enorme impacto no processo de descoberta dos novos conhecimentos produzidos. A internet facilitou o compartilhamento das informações, mas seu tamanho tende a dificultar todo esse procedimento. Neste cenário, a produção científica vem ganhando espaço, fazendo surgir a necessidade de um processo de qualificação e análise mais profunda dos dados disponíveis. Base de dados com essas informações surgem para facilitar o acesso às pesquisas mais recentes e disponibilizam informações relevantes. Porém, há muito trabalho a ser realizado para que essas informações sejam cada vez mais relevantes aos pesquisadores e órgãos de fomento à pesquisa. Por isso é que nesse contexto, sistemas de recomendação tornam-se promissores pela capacidade que possuem para ajudar a resolver a sobrecarga de informação, identificando de forma personalizada, dentro de um contexto, as informações relevantes, e a personalização do conteúdo baseado em perfis similares se destaca, tanto para recomendar artigos similares, quanto para sugerir novos parceiros para pesquisas. Para tanto, neste trabalho, utilizou-se dos dados do currículo do pesquisador para montar seu perfil, o qual expressa suas preferências através do histórico da vida cientifica, indicando seu futuro acadêmico, suas preferências referente a pesquisas e aponta conexões futuras com outros pesquisadores. Dessa forma, este trabalho tem por objetivo apresentar uma abordagem para identificar a similaridade entre perfis de pesquisadores, que oportuniza a geração de recomendações baseadas em seus perfis. Além disso, buscou-se estruturar um modelo de perfil de pesquisadores para identificar as similaridades entre os perfis, propondo uma métrica para calcular tais similaridades, com o objetivo final de utilizar o modelo de perfil e a métrica de similaridade junto ao mecanismo de recomendação. Ainda, como contribuição do trabalho, foi criada uma ferramenta que utiliza o modelo de perfil, a métrica de similaridade com o mecanismo de recomendação. Para o desenvolvimento da abordagem, utilizou-se conceitos de personalização e sistemas de recomendação. Através de experimentos realizados com dados reais de pesquisadores de oito áreas do conhecimento obtidos junto a Plataforma Lattes, observou-se que a abordagem proposta tende a ajudar a descoberta de conteúdo útil ao pesquisador, através de recomendações. Os experimentos também demonstram que a abordagem proposta tem uma boa cobertura de recomendações. Da mesma forma, através de cálculos junto aos termos minerados no currículo, conseguiu-se apreender as preferências e identificar mudanças nas mesmas, ao analisar o currículo do pesquisador junto a aspectos temporais. Os indicadores de similaridade não só apontam semelhanças entre os perfis, como também ajudam a geração das recomendações.

Palavras-chave: perfil de pesquisadores, similaridade de perfil, sistemas de recomendação.

AN APPROACH TO IDENTIFY SIMILARITY BETWEEN RESEARCHER PROFILES WITH A VIEW TO THE RECOMMENDATION

ABSTRACT

The production of knowledge by humanity increases with each passing day, generating a huge impact in the process of discovering the new knowledge produced. The internet has made it easier to share information, but its size now tends to make this whole thing difficult. In this scenario, scientific production has been gaining ground, raising the need for a process of qualification and deeper analysis of available data. Database with this information appears to facilitate access to the latest searches and provide relevant information. However, there is much work to be done to make this information increasingly relevant to researchers and research promotion agencies. That is why in this context, recommendation systems become promising for their ability to help resolve information overload by identifying, in a context-specific, context-sensitive information, and customizing content based on similar profiles stands out, both for recommending similar articles and for suggesting new partners for research. To do so, we used the data from the researcher's curriculum to build his profile, which expresses his preferences through the history of scientific life, indicating his academic future, his preferences regarding research and points future connections with other researchers. Thus, this paper aims to present an approach to identify the similarity between profiles of researchers, which allows the generation of recommendations based on their profiles. In addition, we tried to structure a profile model of researchers to identify similarities between the profiles, proposing a metric to calculate such similarities, with the final objective of using the profile model and the similarity metric next to the recommendation mechanism. And, create a tool that uses the profile model, the similarity metric with the recommendation engine. For the development of the approach, we used personalization concepts and recommendation systems. Through experiments carried out with real data from researchers from 8 areas obtained from the Lattes Platform, it was observed that the proposed approach tends to help the discovery of useful content to the researcher through recommendations. Experiments also demonstrate that the proposed approach has a good coverage of recommendations. In the same way that, through calculations with the terms mined in the curriculum, it was possible to apprehend the preferences and identify changes in them, when analyzing the curriculum of the researcher along temporal aspects. Not only do similarity indicators point to similarities between profiles, but they also help to generate recommendations.

Keywords: researcher profile, profile similarity, recommender systems.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ACM – Association for Computing Machinery

CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico

DBLP – Digital Bibliography e Library Project

DTD – Document Type Definition

IEEE – Institute of Electrical and Electronics Engineers

TCC – Trabalho de Conclusão de Curso

TF-IDF – Term Frequency – Inverse Document Frequency

XML – eXtensible Markup Language

ETL – Extract Transform Load (Extração, Transformação e Carga)

SUMÁRIO

1. INTRODUÇÃO

Na atualidade, estudos sobre a produção científica de pesquisadores vêm ganhando espaço, principalmente, pela necessidade de um processo de qualificação da gestão dos recursos aplicados junto ao financiamento de pesquisas. Nesse contexto, a análise de base de dados de produção científica é fundamental para a tomada de decisões pelos órgãos de fomento, já que influencia diretamente na decisão de onde investir mais recursos. Assim, é fato que conhecer os pesquisadores e suas atividades pode contribuir efetivamente para uma correta avaliação na aplicação dos recursos, da mesma forma que, essa análise de perfil, pode guiar as avaliações para a contratação e/ou promoção.

Nesse sentido, como afirmam Cervi, Galante e Oliveira [1], "A tarefa de avaliar a produção científica de um pesquisador é baseada fortemente na análise dos dados disponíveis em seu currículo." Assim, o currículo torna-se elemento fundamental, já que contém todo o histórico do pesquisador, expressando todo o seu perfil, podendo indicar muito sobre seu futuro e as preferências referente as suas pesquisas e conexões com outros pesquisadores.

No Brasil, os pesquisadores tem a possibilidade de informar suas atividades junto a plataforma Lattes[1], uma iniciativa relevante para arquivamento das atividades acadêmicas. A plataforma, segundo Lima *et al.* [2], tem por objetivo ser um repositório das atividades desenvolvidas por pesquisadores e fornece uma padronização para a publicação do currículo do pesquisador. Outra iniciativa de destaque é a plataforma DBLP[2], que consiste em um repositório bibliográfico de ciência da computação, muito utilizado para recuperar detalhes bibliográficos ao compor referências. Trabalhos com de Lima *et al.* [2], utilizam tanto a lista de títulos de periódicos do Lattes quanto da DBLP.

Neste contexto, Cervi, Galante e Oliveira [1] avaliam que a plataforma Lattes, apesar de ser um importante banco de dados referente às produções científicas de pesquisadores brasileiros, não possui uma consulta sofisticada sobre os dados. Assim, o processo acaba por ser realizado manualmente, demandando tempo, sendo exaustivo e podendo levar a situações equivocadas, tendo em vista que a análise humana perpassa uma série de fatores e, em análise de um grande volume de dados, pode ser falha.

De acordo com Lima *et al.* [2], a avaliação de pesquisadores é exaustiva e necessita de comissões altamente especializadas, sendo que a verificação de currículos

[1] Plataforma Lattes. Disponível em http://lattes.cnpq.br.
[2] Plataforma DBLP (Digital Bibliography & Library Project). Disponível em http://dblp.uni-trier.de.

acadêmicos realizada manualmente não é somente trabalhosa, mas pode gerar erros de avaliação sendo injusto com os avaliados. Outro aspecto também reside na necessidade de verificação do perfil como um todo e não somente publicações para a avaliação de pesquisadores e, ainda, há de se considerar as peculiaridades de diferentes áreas.

Ressalta-se também que esta dificuldade de análise mais completa do perfil do pesquisador vem associada a alguns desafios, tais como: (i) onde buscar os dados e quais atributos utilizar para comparar; (ii) como identificar comportamentos semelhantes; (iii) quais métricas utilizar para avaliação, aspectos que, se realizados de forma manual, geram um trabalho moroso e exaustivo, passível de vário equívocos.

Da mesma forma, outra dificuldade que se apresenta ao pesquisador é que o próprio conhecimento descoberto é amplamente divulgado em formato digital e arquivado em todo o mundo, muitas vezes de forma pouco contextualizada. O pesquisador moderno tem nível sem precedentes de acesso à soma total do conhecimento humano. Isso certamente cria um problema conhecido como "sobrecarga de informação". Nesse contexto, os pesquisadores encontram um número avassalador de material para suas consultas de pesquisa, mas para os quais a maioria são em grande parte irrelevantes para as suas necessidades [3].

Assim, é latente o surgimento de uma crescente necessidade de estudos que permitam a criação de ferramentas que ofereçam a possibilidade de cruzamento de dados dos pesquisadores. A realização de tais estudos pode proporcionar a análise da rede de coautoria dos pesquisadores, o nível de integração entre eles, as possibilidades de colaboração científica na produção conjunta, bem como orientações de mestrado e de doutorado em parceria. Tais situações, poderiam gerar recomendações com base no perfil similar dos pesquisadores, o que seria útil não somente às agências de fomento ou instituições de pesquisa, mas também para os próprios pesquisadores. Nesse contexto, pode-se afirmar que sistemas de recomendação são uma abordagem promissora para resolver a sobrecarga de informação, levando-se em conta sempre o perfil de cada pesquisador para a recomendação de materiais relevantes ao seu contexto. Como consideram Hannel *et al.* [4], um perfil bem definido pode gerar uma base de um sistema de recomendação relevante, de artigos científicos para os pesquisadores.

A interação entre os pesquisadores, tem sido utilizada de forma constante em diversas áreas, tais como redes de colaboração científica e identificação de pesquisadores com perfis semelhantes. Já a similaridade para sistemas de recomendação ganha força para recomendar obras similares ou sugerir novos parceiros, e as redes de colaboração científica estão mudando a maneira de trabalhar com pesquisa e inovação. Identificando pesquisadores com perfis semelhantes há o estímulo ao trabalho colaborativo para melhorar a qualidade das

publicações científicas, da mesma forma que melhorar o intercâmbio entre pesquisadores aumenta a troca de experiências entre grupos de pesquisa e promove a expansão de redes de colaboração científica [5].

De acordo com Lee, Lee e Kim [6], a quantidade de trabalhos acadêmicos publicados em revistas e conferências só tende a aumentar com o passar dos anos, e cada vez mais fica difícil para os pesquisadores descobrir novos artigos que dizem respeito ao seu trabalho, fazendo a busca simplesmente através da verificação das revistas da área em que atua ou através de palavras chaves em buscadores. No processo conduzido dessa forma, cada vez mais o pesquisador gasta um tempo importante para atualizar seus conhecimentos relacionados ao seu campo de pesquisa. É por esse motivo que, neste contexto, os sistemas de recomendações cada vez mais personalizados, tendem a ajudar no ganho de tempo ao recomendar trabalhos mais relevantes para o pesquisador.

Igualmente, é fato que hoje o pesquisador carece de ferramentas capazes de responder de forma proativa as dificuldades cada vez maiores em relação ao grande volume de dados que está sendo produzido pela humanidade. Tanto a plataforma Lattes quanto a DBLP, apesar de terem filtros rápidos em suas buscas, ainda necessitam de mecanismos que melhorem a exploração junto aos seus dados. Portanto, o desenvolvimento de uma abordagem que permita a melhor exploração dos dados destas plataformas, onde os mesmos possam ser correlacionados de formas diferentes, claramente resultará em ganho de tempo e de uma assertividade maior junto a decisões tomadas na análise dos dados e para o desenvolvimento de novas pesquisas e contatos com outros pesquisadores.

Foi com base em toda essa problemática prática e teórica que vem sendo examinada pelos estudiosos do tema, que então surge o problema de pesquisa do presente estudo: como identificar que pesquisadores possuem perfis similares para que possam ser geradas recomendações baseadas em seus perfis?

Como se pode observar no trabalho de Gollapalli, Mitra e Giles [7] é abordado um sistema de recomendação de pesquisadores com instâncias similares para o domínio acadêmico, propondo modelos de semelhanças entre pesquisadores com base nos dados extraídos de suas publicações e páginas pessoais. Os resultados apontam que as técnicas de representação do perfil do pesquisador necessitam ser melhoradas, contendo mais informações e metadados. São trabalhos como este que demonstraram que existem muitos desafios a serem explorados no tema que envolve similaridade de perfil de pesquisador e recomendação acadêmica. Diante disso, cria-se um espaço importante para que soluções tecnológicas sejam implementadas a fim de minimizar ou até mesmo acabar com essas

dificuldades, o que pode ser aplicado em diversas áreas de pesquisa, cujos exemplos podem ser as aplicações para dispositivos móveis, adaptabilidade frente ao contexto, recomendação de informações desejadas pelo usuário.

Em trabalhos como de Hannel *et al.* [4], Mena-Chalco, Digiampietri e Cesar-Jr [8] e Hong, Jeon e Jeon [9], a similaridade é utilizada para a desambiguação dos dados bibliográficos ou para a verificação da semelhança entre *strings* retirados dos textos dos trabalhos dos pesquisadores, para verificar o quão semelhantes são e, assim, sugerir estes a outros pesquisadores utilizando-se, assim, apenas um atributo do perfil dos pesquisadores, suas produções, e não com a amplitude de seu currículo o qual descreve sua vida profissional. Dessa forma, as sugestões parecem ser bastante centradas em relação aos trabalhos bibliográficos e não em outros aspectos que compõem a totalidade do perfil do pesquisador. Igualmente, nesse processo, não se trabalha a questão envolvendo a ponderação dos atributos referentes à similaridade do perfil do pesquisador como também, não se busca uma personalização do conteúdo estruturada junto ao perfil do pesquisador.

Pelo que foi visto, a literatura apresenta diferentes técnicas para medir a similaridade. No entanto, por mais que estes métodos tenham sidos bem sucedidos em casos específicos onde foram utilizados, não foi possível detectar uma forma padronizada para avaliar a similaridade de um perfil de pesquisador. Nesse contexto, a extração de característica e medida de similaridade é um desafio, devido a uma grande variedade de funções de medida de similaridade que podem ser combinadas com as diferentes técnicas que retornam resultados que nem sempre são os mais satisfatórios. Busca-se aqui, uma nova métrica para calcular a similaridade do perfil do pesquisador, usando dados de seu currículo, não levando em consideração somente aspectos de suas publicações, com uma visão mais abrangente e adaptável aos mais diversos contextos.

Focado nos desafios apresentados, o presente trabalho visa desenvolver uma abordagem para identificar a similaridade entre perfis de pesquisadores e a geração de recomendação baseada nestes perfis. Para atender aos objetivos, passos específicos foram determinados a serem realizados: (i) estudar dados de pesquisadores oriundos de produção científica; (ii) estruturar um modelo de perfil de pesquisadores baseados na similaridade de perfis; (iii) propor uma métrica para calcular a similaridade entre os perfis dos pesquisadores; (iv) definir um mecanismo de recomendação que utilize o modelo de perfil elaborado e a métrica proposta; (v) desenvolver uma ferramenta com a métrica e com o mecanismo de recomendação.

A fim de descrever a abordagem proposta, a ferramenta web desenvolvida e os experimentos, o presente trabalho encontra-se organizado conforme segue. O Capítulo 2 apresenta os principais conceitos que fundamentam esta dissertação, bem como os trabalhos relacionados pesquisados na literatura. No Capítulo 3 a abordagem proposta é especificada, detalhando a modelagem do perfil, a similaridade, o processo de coleta dos dados, as métricas, o sistema de recomendação e o desenvolvimento da ferramenta web. Na sequencia o Capítulo 4 descreve os experimentos e a análise dos resultados obtidos. Por fim, o Capítulo 5 apresenta as considerações finais e trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

O presente Capítulo tem o objetivo de apresentar o contexto sobre perfil de pesquisadores, similaridade de perfil de pesquisadores e sistemas de recomendação acadêmica. Na Seção 2.1 são apresentados conceitos sobre perfil de usuário e modelagem de perfil. Já na Seção 2.2, são apresentados o contexto sobre perfil de pesquisadores, o currículo e sua produção científica como também é explanado sobre bases de dados de produção científicas, as quais tenham seus dados em formato aberto para utilização e, na sequência, a Seção 2.3 expõe uma visão geral sobre similaridade em perfil. As métricas no contexto de similaridade de perfil de pesquisadores são demonstradas na Seção 2.4 e os conceitos sobre sistemas de recomendações e sistema de recomendação no contexto de pesquisadores estão na Seção 2.5. Nas Seções 2.6 e 2.7, são analisadas, brevemente, as redes sociais e redes de coautoria acadêmicas. Já na Seção 2.8 apresenta-se os trabalhos analisados que possuem relação com os temas similaridade de perfil de pesquisadores e recomendação acadêmica, com o objetivo de compará-los, demonstrando os pontos em aberto entre cada um deles. Por fim, a Seção 2.9 apresenta as considerações finais do Capítulo.

2.1. PERFIL DE USUÁRIOS

Segundo os dicionários Aurélio[3] e Michaelis[4], perfil é um conjunto de características ou competências. Trata-se de relato breve em que se apresenta a vida de uma pessoa, podendo ser pessoal, profissional ou referente a uma atividade específica. O perfil é diferente para cada indivíduo porque, embora as categorias sejam as mesmas para cada conceito, este será fortemente influenciado pelas experiências distintas de cada um. Porém, eles poderão ser altamente semelhantes, uma vez que o perfil de usuário faz parte de uma classe, contendo atributos que não serão tão diferentes da normalidade das características ou competências de determinado conceito.

A representação dos interesses do usuário, e não somente suas características e seus dados, pode ser modelada através da definição de seu perfil. A evolução de um perfil deve ser levada em consideração, uma vez que o conhecimento e o interesse do usuário são dinâmicos, transformando-se com o passar do tempo. Neste contexto, é necessário manter um

[3] Disponível em https://dicionariodoaurelio.com.
[4] Disponível em http://michaelis.uol.com.br.

histórico dos interesses dos usuários, ou seja, acompanhar a trajetória do perfil, para que um sistema represente ao máximo os interesses do usuário. O perfil de usuário é a base para a personalização de um sistema de recomendação, pois representa o comportamento e o conhecimento desse usuário.

O perfil de usuários está relacionado com personalização e recomendação, sendo um tema largamente discutido em vários artigos científicos. Para Cervi, Galante e Oliveira [1], as diversas abordagens, métodos e sistemas existentes se aplicam em diversas áreas, como por exemplo: (i) detecção de perfil de clientes em sites de comércio eletrônico; (ii) recomendação de produtos; (iii) recomendação de especialistas para participarem de defesas de mestrado e de doutorado; (iv) recomendação de pesquisadores para formação de rede de colaboração científica; (v) organização de comitê de programa de conferência científica; (vi) organização de corpo editorial e revisores de periódicos; (vii) medição do impacto de publicações, entre outras.

Dentro deste contexto, a área de modelagem de perfil tem como objetivo descobrir conhecimento sobre o usuário em determinado assunto e como representar esse conhecimento [10]. Diversos sistemas e métodos tem sido propostos para se modelar o perfil de um usuário, sendo de forma explícita, implícita ou ainda, da fusão das duas técnicas [11]. A modelagem do perfil de um usuário pode ser baseada em seu conhecimento ou em seu comportamento [12]. Nesse sentido, a modelagem de um usuário envolve compreender informações sobre o usuário através de informações observadas sobre o mesmo, como suas ações, seu estado, seu comportamento ou seu conhecimento. Já na modelagem baseada em conhecimento, comumente os usuários são associados a modelos estáticos de usuários e o processo de definição do perfil é conduzido por entrevistas ou questionários. Neste, o usuário deve participar de forma ativa do processo, uma vez que as informações relevantes para a definição do perfil devem ser fornecidas pelo mesmo. De outra forma, na modelagem baseada em comportamento, o princípio que norteia o modelo é o próprio comportamento do usuário. Assim, o mesmo, não participa do processo de modelagem de forma explícita, pois a definição do perfil pode ser dar tanto por meio de técnicas de mineração de dados, quanto por aprendizagem de máquina para descobrir padrões úteis de seu comportamento.

O processo de modelagem do perfil de usuário envolve a identificação de quais dados são relevantes para a definição de um perfil e, pode ser apresentada por meio de uma ontologia, de um vetor de termos representados por palavras-chave, de um arquivo XML, dentre diversas outras formas. Ontologia é uma especificação explícita de uma conceitualização, onde define-se os termos usados para descrever e representar uma área do

conhecimento [13]. Deste modo, ontologias são usadas por pessoas e aplicações para a troca de informações sobre um determinado domínio, como também fornecem definições de conceitos básicos de um domínio apropriados para o processamento automático. Outra forma de representação de perfil de usuários é por meio de vetores de termos, onde comumente é definido por palavras-chave que representam as características do usuário para a composição de seu perfil.

Para Trajkova e Gauch [10] a base sólida para a construção do modelo é que este represente exatamente os interesses do usuário, independentemente do ambiente, através de três objetivos principais: (i) descobrir o conhecimento ou interesse de um usuário em determinado assunto; (ii) representar e armazenar este conhecimento ou interesse internamente em um sistema; (iii) gerenciar alterações no conhecimento ou interesse. Levando-se em consideração que os interesses de um usuário tendem a mudar com o tempo, a representação do perfil deve refletir possíveis mudanças, uma vez que é necessário representar com o máximo de exatidão os interesses e preferências do usuário.

Normalmente, apenas as características mais descritivas são utilizadas para modelar um item e/ou usuário, uma vez que os atributos mais discriminativos são identificados e são armazenados, frequentemente, como um vetor que contém as características e os seus pesos. O modelo de perfil normalmente consiste nas características de itens de um usuário e para gerar as recomendações, o modelo de perfil e os candidatos à recomendação são comparados [14].

Dentro desse enfoque, Montaner [15] aborda questões que vão além das definições do perfil, apresentando a necessidade de pensar nas atualizações do mesmo. Assim, aponta cinco etapas para a modelagem do perfil: (i) uma técnica de representação do perfil; (ii) uma técnica usada para gerar o perfil inicial; (iii) *relevance feedback* que represente os interesses do usuário; (iv) uma técnica de aprendizagem de perfil; e (v) uma técnica de adaptação de perfil.

Já no trabalho desenvolvido por Syed e Andritsos [16], analisou-se a modelagem das preferências do usuário, apresentando um comparativo de trabalhos na área de filtragem de informação e sistemas de recomendação. Foram utilizados três critérios de abordagens, quais sejam, o perfil do usuário, esquema de filtragem e métricas de avaliação. No perfil, aborda-se como o sistema arquiteta e armazena o perfil do usuário, bem como os métodos usados para capturar os interesses do usuário. No esquema de filtragem expõe a função de classificação ou algoritmo usado pelo sistema. Já nas métricas de avaliação verifica a sua existência, tendo, então analisa a métrica utilizada.

2.2. PERFIL DE PESQUISADORES

A gestão da produção científica, atualmente, busca processos de qualificação para decidir onde aplicar recursos implicando, claramente, que o fomento à pesquisa seja cada vez mais eficiente. Assim, conhecer o perfil do pesquisador é, cada vez mais, atividade de suma importância para os órgãos de fomento. Na literatura, existem vários trabalhos abordando questões de produção cientifica de pesquisadores, avaliando a produção científica e realizando estudos comparativos [17], [18]. Como Wainer e Vieira [19] correlacionam em seu trabalho, no qual demonstram avaliações realizadas por comissões disciplinares com algumas medidas de dados bibliométricos, a fim de otimizar a analise curricular dos pesquisadores, como também apresentam a comparação do perfil influenciando as decisões de aumentar, manter ou diminuir bolsas de pesquisas financiadas por órgãos de fomentos.

Da mesma forma que, são analisadas decisões sobre os pesquisadores brasileiros e computado suas relações com 21 medidas diferentes, entre elas: produção total, produção dos últimos 5 anos, total de citações recebidas entre outras. Aqui cabe ressaltar, que as métricas para avaliação da produtividade, baseadas em produção (número total de artigos escritos, número total de artigos indexados), produtividade e impacto do trabalho do pesquisador são ponto de destaque importantes. Assim, a produtividade é avaliada por medidas de produção calculadas ao longo de um intervalo de tempo, por exemplo, o número total de artigos indexados nos últimos 5 anos. Já as medidas de impacto normalmente referem-se às citações recebidas pelos artigos publicados pelo pesquisador, tais como, o número total de citações recebidas, número médio de citações por ano, e assim por diante. Da mesma forma que, Wainer e Vieira [19] enfatizam claramente em seu trabalho que o componente mais importante na avaliação de um pesquisador é o curriculum vitae.

No contexto de perfil de pesquisadores, Hannel, Warpechowski e Lima [20] apresentam que perfil é um conjunto de características e critérios de qualidade que identificam os pesquisadores. As características são informações pessoais como idiomas, endereço, dentre outros e os critérios de qualidade são informações quantitativas e qualitativas.

Já no trabalho de Hong, Jeon e Jeon [9], foi proposto um sistema de recomendação de artigos baseado no perfil do usuário, onde é feito a extração da palavras-chave dos artigos pesquisados pelo pesquisador para compor um perfil deste pesquisador. Se as palavras-chave do artigo estiverem inacessíveis por algum motivo, o título do trabalho gera uma combinação de palavras-chave que são associadas ao perfil do usuário. A relevância da informação recomendada está relacionada com as preferências do perfil do usuário. Neste

contexto, o perfil de usuário é geralmente representado como um conjunto de palavras-chave ponderado, redes semânticas ou regras de associação e é construído a partir de fontes de informação usando uma variedade de técnicas de construção com base na aprendizagem de máquina ou recuperação de informação. O perfil é armazenado na forma de um arquivo XML, e atualizado a cada clique dado pelo usuário em determinado trabalho de pesquisa, recalculando a taxa de ocorrência e refletindo no perfil do usuário.

Nesse mesmo caminho, Cervi, Galante e Oliveira [5], [11] identificaram a reputação de pesquisadores usando um modelo de perfil adaptativo denominado Rep-Model. No seu trabalho, em primeiro lugar, definiu-se aonde encontrar e quais dados são utilizados para avaliar a reputação de um pesquisador e ainda, como representar estes dados em um modelo de perfil. Em seguida, definiu-se uma métrica para chegar a um índice da reputação, a qual chamaram de Rep-Index. O modelo de perfil e a métrica foram definidos baseados no equilíbrio da trajetória do perfil de pesquisador, abrangendo os principais elementos de seu currículo. Tanto o perfil quanto a métrica não focaram apenas em publicação de artigos ou citação, mas na inclusão da vida científica do pesquisador ao longo de sua carreira.

O modelo de perfil de pesquisador também é trabalhado por Hannel, Warpechowski e Lima [20] para medir a qualificação dos pesquisadores. O modelo foi definido através de uma ontologia denominada OntoResearcher. No contexto do trabalho, a ontologia é definida como a representação de termos, definições e critérios de qualidade do conceito pesquisador.

Alguns trabalhos abordam a identificação do perfil de pesquisadores, através da extração de informações da web para modelagem de seu perfil, enquanto outros ainda fazem a identificação baseados em dados científicos ou simplesmente análise de produção para obter estatísticas relevantes. Nesse sentido, é possível destacar Vivian e Cervi [21], que em seu trabalho utilizam ferramentas de recuperação de informações de dados através da web, para compor o perfil do pesquisador através de plataforma específica. No trabalho dos autores, utilizou-se dados que nada mais são que as informações do currículo do pesquisador, como formação acadêmica, produção acadêmica (artigos, livros, capítulos de livros, produção bibliográfica, trabalhos em congressos, trabalhos técnicos e resumos expandidos) e orientações (TCC, mestrado, doutorado e iniciação científica). Esses e outros dados compõem um *dataset* para a comparação dos perfis dos pesquisadores o que permite identificar a reputação através do conjunto de indicadores.

Já no trabalho de Sugiyama e Kan [3], o perfil de usuário de cada pesquisador é criado utilizando as respectivas listas de publicações da plataforma DBLP. Estes autores

utilizaram-se das citações para montar as preferências do perfil, assim podendo gerar recomendações a estes. De outra forma, Nascimento *et al.* [22] utilizaram-se do título e do resumo dos trabalhos para a construção do perfil de usuário, gerando *tags* das características encontradas nos textos.

Para Cervi, Galante e Oliveira [5], [11], no contexto acadêmico, a identificação do perfil pode ocorrer por meio de um processo de análise da trajetória da carreira de um pesquisador. Para esses autores, o processo envolve não somente aspectos relacionados a produção científica, mas também por outros elementos essenciais a atividade de um pesquisador, ou seja, utilizando a informação de todo o seu currículo.

2.2.1. Plataforma Lattes

O Lattes é um banco de dados de currículos, grupos de pesquisa e instituições, sendo uma plataforma virtual criada e mantida pelo CNPq[5] com o intuito de facilitar as ações de planejamento, gestão e operacionalização do fomento à pesquisa. O currículo Lattes tornou-se um padrão nacional no registro da vida pregressa e atual de estudantes e pesquisadores do país, sendo adotado pelas instituições de fomento, universidades e institutos de pesquisa, tornando-se assim, um elemento indispensável à análise de mérito e competência dos pleitos de financiamentos na área de ciência e tecnologia. Como os dados são públicos, há uma maior transparência e confiabilidade às atividades de fomento do CNPq e das agências que o utilizam, além do fato de que os mesmos fortalecem o intercâmbio entre pesquisadores e instituições e são uma fonte de informações para estudos e pesquisas.

A Figura 1 apresenta uma estatística da base de dados que a plataforma Lattes possui. O perfil do currículo é bem amplo, contendo produção acadêmica como artigos, livros, capítulos de livros, produção bibliográfica, trabalhos em congressos, trabalhos técnicos e resumos expandidos. Há também informações acerca de orientações de TCC (Trabalho de Conclusão de Curso), mestrado, doutorado e iniciação científica.

[5] Conselho Nacional de Desenvolvimento Científico e Tecnológico. Disponível em http://www.cnpq.br.

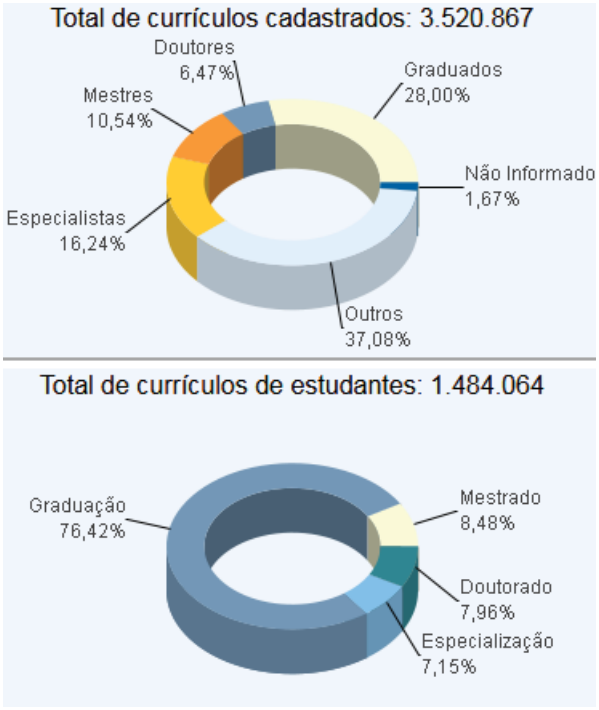


Figura 1: Estatísticas da Base de Currículos da Plataforma Lattes[6].

Essa base de dados é bastante utilizada em diversos trabalhos, como é o caso de Mena-Chalco, Digiampietri e Cesar-Jr [8], onde os dados do Lattes são utilizados para a caracterização de uma rede de coautoria e é avaliado possíveis inconsistências dos dados, devido a erros de digitação ou falta de padronização na escrita dos nomes dos coautores. Assim, a comparação de duas produções é realizada pelo casamento aproximado entre os títulos associados a cada produção e os títulos são considerados iguais ou equivalentes se ambos forem similares em 90%. A similaridade entre duas *strings* baseia-se na distância proposta por Levenshtein, obtida pelo número mínimo de inserções, eliminações ou substituições de caracteres necessários para transformar um texto em outro. A plataforma Lattes é uma base de dados pública, tanto no que se refere ao ingresso e cadastramento dos currículos, quanto ao acesso das informações destes. Apesar da disponibilização pública das informações, atualmente, têm-se solicitado ao CNPq o acesso aos dados brutos, com o objetivo da realização de estudos de bibliometria, cientometria, entre outros, como também para gerar indicadores internos de produção científica e tecnológica. A extração de dados está disponível a todas as instituições de ensino e pesquisa e inovação do país, por meio de um acordo institucional, contendo a exposição de motivos e destinação a ser dada aos dados a serem extraídos. Os dados são disponibilizados no formato padrão XML (eXtensible Markup

[6] Fonte: http://estatico.cnpq.br/painelLattes (Extração dos dados em 31/11/2016).

Language), Figura 2, definido junto à Comunidade Conscientias-LMPL[7], responsável pela padronização das informações existentes na plataforma Lattes.

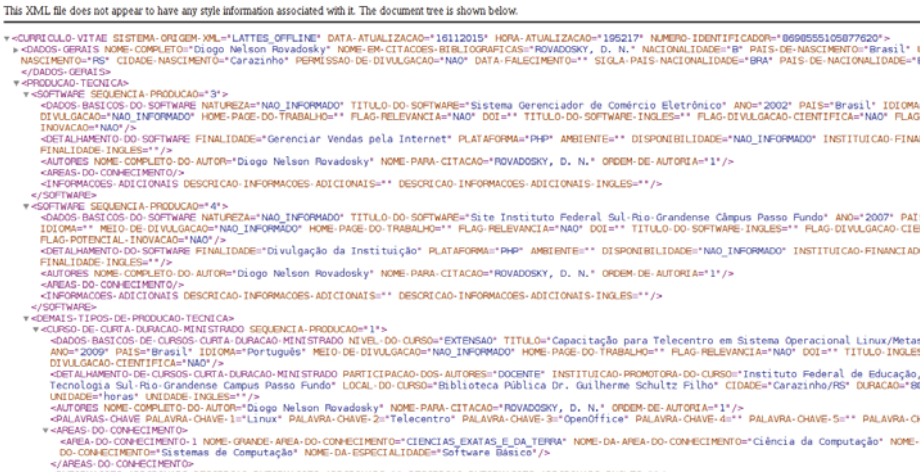


Figura 2: Visualização de parte de um currículo em arquivo XML.

O padrão que define os tipos de informação existentes em um documento XML é dado pelo seu respectivo DTD (Document Type Definition). Essa linguagem para modelagem de arquivos XML permite que aplicativos compreendam a estrutura em uma árvore definida no documento XML. A Figura 3 mostra parte do arquivo DTD mantido pela Comunidade Conscientias-LMPL.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!--
##########################################################################################
#################    -->
<!-- Ontologia XML para a Unidade de Informação de Currículos. -->
<!-- Todos os direitos reservados a Comunidade CONSCIENTIAS/LMPL
(http://lattes.cnpq.br/lmpl). Versão 1.4.1 - Atualizada em 03/03/2004. -->
<!-- Este documento não deve ser alterado. -->
<!-- Dúvidas e sugestões podem ser enviadas a lmpl@cnpq.br. -->
<!-- O histórico das versões desta ontologia podem ser obtidas em:
http://www.cnpq.br/lmpl/documentacao/curriculo/historico.htm -->
<!-- Versão 1.1 (Atualizada em 28/03/2003) Alteração necessária devido a alteração ocorrida
no Lattes 1.5. Foi-se inserido o valor "RESUMO_EXPANDIDO" no atributo "NATUREZA" do elemento
"DADOS-BASICOS-DO-TRABALHO".  -->
<!-- Versão 1.2 (Atualizada em 16/04/2003) Correção do tamanho máximo de caracteres dos
atributos do tipo Memo no Lattes, corrigido de 64000 para 2000.  -->
<!-- Versão 1.2.1 (Atualizada em 24/04/2003) Alteração no tipo MES, a partir desta versão
estará sendo utilizado o tipo gMonth do XML Schema. -->
<!-- Versão 1.3 (Atualizada em 12/05/2003) Alterações: Inclusao da opção "NAO_INFORMADO" nos
atributos: PROFICIENCIAS de IDIOMAS, e TIPO-DE-EVENTO de
DETALHAMENTO-DA-APRESENTACAO-DE-OBRA-ARTISTICA e DETALHAMENTO-DA-OBRA-DE-ARTES-VISUAIS. -->
<!--Versão 1.3.1 (Atualizada em 20/08/2003) Alteração: Mudança do tipo NMTOKENS para CDATA
para permitir o caracter apóstrofo no atributo NOME-COMPLETO. Segue nova especificação para
esse: "<!ATTLIST DADOS-GERAIS NOME-COMPLETO CDATA #REQUIRED" (somente no DTD).-->
```

Figura 3: Visualização de parte do arquivo DTD (LMPLCurriculo.dtd).

Muitos trabalhos utilizam-se de ferramentas próprias para a extração dos dados da plataforma Lattes. Mena-Chalco e Junior [23] descrevem um sistema de código aberto que

[7] http://lmpl.cnpq.br/lmpl.

automaticamente utiliza os dados da plataforma Lattes. Esse código denominado scriptLattes[8], é utilizado em trabalhos como o de Magalhães *et al.* [24], o qual tem o objetivo de identificar e extrair a produção científica, produtos tecnológicos, instituições e rede de cientistas que trabalham com um certo tipo de doença. Já no trabalho de Giordano, Bruning e Bordin [25], o scriptLattes é utilizado para analisar uma rede de colaboração científica. A escolha do scriptLattes para o trabalho citado, possibilitou o acesso às informações dos pesquisadores cadastrados na plataforma Lattes a partir do número identificador de cada currículo.

2.2.2. Plataforma DBLP

A DBLP é um repositório bibliográfico de ciência da computação hospedado na Universidade Trier, na Alemanha, consolidado em um serviço aberto que fornece informações bibliográficas de importantes periódicos e conferências de ciência da computação. Em junho de 2009, o DBLP continha mais de 1,2 milhões de registros bibliográficos, segundo Ley [26]. Nesse sentido, a Figura 4 mostra o número de novos registros no banco de dados do site DBLP, por tipo de publicação e ano. Para os pesquisadores da área de ciência da computação o site DBLP é uma ferramenta popular para rastrear o trabalho de colegas e para recuperar detalhes bibliográficos ao compor suas referências. Os dados do DBLP podem ser baixados e os registros bibliográficos estão contidos em um arquivo XML.

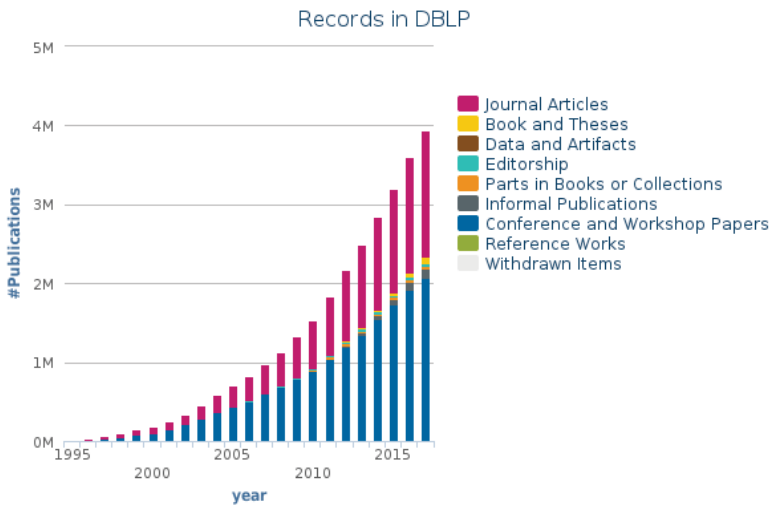


Figura 4: DBLP novos registros por ano e tipo de publicação[9].

[8] http://scriptlattes.sourceforge.net
[9] Fonte: http://dblp.uni-trier.de/statistics/recordsindblp.html.

De acordo com Ley [26], as principais vantagens da DBLP são a livre disponibilidade e a inclusão de muitos anais de congressos, que desempenham um papel essencial para muitos ramos da ciência da computação e que são mal cobertos por outras bases de dados bibliográficas. Outra vantagem é a facilidade da geração de diversos gráficos/grafos a partir dos dados como, por exemplo, o grafo de coautoria, que é um bom exemplo para redes sociais acadêmicas.

Os dados DBLP, além de terem um dos maiores repositórios de citações em ciência da computação, são mantidos com intervenção humana maciça nas fases de aquisição e de carga, com rígidas exigências de qualidade dos dados. Tendo seu conteúdo altamente normalizado em relação ao autor e nomes de publicação, limita a ocorrência de ambiguidade entre os nomes de autores, facilitando a correta identificação e classificação das revistas e conferências. Somado a isso, os dados são totalmente abertos, sem interesse comercial, o que facilita o acesso. Como consequência destes fatores, a utilização dos dados DBLP é largamente difundida como uma das principal fontes de dados para estudos bibliométricos, análise de rede de colaboração e citação [27][28].

Laender *et al.* [29] lembram que, além da sua boa cobertura de conferências, o site DBLP também abrange uma parte substancial de revistas importantes para a área da ciência da computação, como a ACM[10] e a IEEE[11]. Nesse mesmo enfoque, em seu trabalho, Gugel *et al.* [30], desenvolvem uma ferramenta para análise quantitativa da produção cientifica de pesquisadores, com o objetivo de promover consultas e cruzamento de dados sobre informações disponibilizadas pela DBLP, através de arquivo XML. O conjunto de dados DBLP está disponível a partir do endereço: http://dblp.uni-trier.de/xml. O arquivo dblp.xml contém todos os registros bibliográficos e é acompanhado pelo arquivo de definição de dados, dblp.dtd. Segundo Ley [26], o arquivo XML é muito simples, apesar de conter um grande quantidade de informações. Como mostra a Figura 5, o elemento raiz XML <DBLP> contém uma sequência longa de registros bibliográficos.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE dblp SYSTEM "dblp.dtd">
<dblp>
  record 1
  ...
  record n
</dblp>
```

Figura 5: Arquivo dblp.xml representação básica.

[10] Association for Computing Machinery. Disponível em http://www.acm.org.
[11] Institute of Electrical and Electronics Engineers. Disponível em https://www.ieee.org.

A Figura 6 mostra parte do arquivo DTD, que lista vários elementos a serem usados como um registro bibliográfico.

```
<!ELEMENT dblp (article|inproceedings|
  proceedings|book|incollection|
  phdthesis|mastersthesis|www)*>
```

Figura 6: Parte do arquivo dblp.dtd.

Na Figura 7, as marcações correspondem aos tipos de entrada utilizados em BibTEX[12], formato de arquivo utilizado para descrever processos de listas de referências, principalmente em conjunto com os documentos de LaTeX[13]. Esses registros DBLP podem ser entendido como "registros BiBTeX na sintaxe XML".

```
<article key="journals/cacm/Szalay08"
     mdate="2008-11-03">
  <author>Alexander S. Szalay</author>
  <title>Jim Gray, astronomer.</title>
  <pages>58-65</pages>
  <year>2008</year>
  <volume>51</volume>
  <journal>Commun. ACM</journal>
  <number>11</number>
  <ee>http://doi.acm.org/10.1145/
     1400214.1400231</ee>
  <url>db/journals/cacm/
     cacm51.html#Szalay08</url>
</article>
```

Figura 7: Parte do arquivo dblp.xml com registros BiBTeX.

Os dados da DBLP são uma fonte importante no que diz respeito a pesquisadores de ciência da computação. No trabalho de Sugiyama e Kan [3], são avaliados publicações de 50 pesquisadores utilizando esses dados, envolvendo vários campos da ciência da computação tais como banco de dados, sistemas embarcados, gráficos, recuperação de informação, redes, sistemas operacionais, linguagens de programação, engenharia de software, segurança e interface de usuários. Nos dados utilizados, não foi encontrado ambiguidade na lista dos 50

[12] Disponível em http://www.bibtex.org.
[13] LaTeX é um sistema tipográfico de alta qualidade; ele inclui funcionalidades concebidas para a produção de documentação técnica e científica. Disponível em https://www.latex-project.org.

pesquisadores, como também verificou-se que os dados são representativos dos principais interesses dos pesquisadores.

2.3. SIMILARIDADE DE PERFIL

Similaridade, segundo dicionário Aurélio[14], é a característica, estado ou natureza do que é similar, ou seja, a particularidade de objetos similares. Similar é o que tem semelhança ou analogia com algo e que permite estabelecer comparações entre duas coisas. A similaridade acontece pelas suas características físicas ou abstratas e, como exemplo, é possível citar dois carros, que podem ser similares por terem linhas em seus *designs* que são parecidas, o que pode confundir num primeiro olhar. Isso significa que a aparência de ambos os carros é similar. A similaridade depende muito da meta de uma aplicação específica. Exemplo, dois carros são similares quando a velocidade máxima é similar? Ou quando o preço é similar? Ou quando a aparência é similar? [31].

Wangenheim, Wangenheim e Rateke [31] apresentam que a similaridade é entendida como sendo a correspondência ou coocorrência de atributos ou características. O conceito de similaridade em um domínio específico pode ser determinado pela Engenharia do Conhecimento e ao se definir conceitos de similaridade que determinam se um caso anterior é similar a uma questão atual, envolveram as seguintes questões: (i) a definição do cenário e das respectivas metas a serem atingidas; (ii) a identificação das entidades de informação importantes para a determinação da similaridade, e de um modelo definindo porque essas entidades são importantes; (iii) a definição de um método para decidir se um caso é similar, definindo um grau numérico de similaridade.

Similaridades são comumente normalizadas em uma faixa de 0 a 1, onde 0 corresponde a dissimilaridade total e 1 a coincidência absoluta. A medida de similaridade é a formalização de uma determinada filosofia de avaliação de semelhança através de um modelo matemático e, possivelmente, a forma mais conhecida de formalização do conceito de similaridade é a definição de uma medida numérica de distância ou similaridade.

No trabalho de Hannel *et al.* [4], foi utilizada como similaridade a função de Smith-Waterman, a qual é elaborada para determinar as regiões similares entre duas sequencias. Em vez de olhar a sequência total, o algoritmo de Smith-Waterman compara segmentos de todos os comprimentos possíveis e otimiza a medida de similaridade. A função

[14] Disponível em https://dicionariodoaurelio.com.

é utilizada após a extração das informações como a do título dos trabalhos de um pesquisador, visto que os dados retirados do currículo podem estar ligeiramente diferente, por erro de digitação ou outro problema qualquer, do que é retornado de outras bases cientificas. Já no trabalho de Hong, Jeoan e Jeon [9] a similaridade é calculada entre um determinado tema e os documentos (artigos) recolhidos usando similaridade do cosseno, com o intuito de recomendar artigos para perfis com temas semelhantes.

Neste contexto, a construção do perfil de um usuário envolve a extração de características do item analisado ao que se deseja associar ao perfil. Em sistemas de recomendação baseados em conteúdo, têm-se usado de itens que contêm informações textuais, como documentos, páginas da web, entre outros. Desta forma, técnicas de recuperação da informação são usadas para extrair palavras-chave deste conteúdo formando assim o perfil, muitas vezes representado por um vetor TF-IDF (*term frequency–inverse document frequency*). A TF-IDF é uma medida estatística que tem o intuito de indicar a importância de uma palavra de um documento em relação a uma coleção de documentos. Assim, o processo baseado em conteúdo elege os itens mais similares aos itens identificados no perfil de interesse do usuário, comparando o perfil do usuário com o perfil de cada item. Entre as técnicas de aprendizagem para detectar e aprender sobre o perfil do usuário e a realimentação de sua relevância estão algoritmos genéticos, redes neurais e classificadores bayesianos. O modelo de espaço vetorial e indexação semântica podem ser empregados por estes métodos de aprendizagem para representar documentos utilizando seus termos ou características. Determinados métodos de aprendizagem, igualmente representam o perfil do usuário com um ou mais vetores, o que torna mais fácil a comparação de documentos e perfis.

Dessa forma, encontrar perfis semelhantes em sistemas de recomendação por filtragem colaborativa é a ideia central, pois segundo os estudos, uma pessoa tende a aceitar a sugestão de um grupo de pessoas próximas ou semelhantes a seu perfil. Assim, a filtragem colaborativa utiliza perfis similares ao do usuário alvo da recomendação, baseado em suas avaliação de um determinado item e sugere este item ao usuário semelhante.

2.4. MÉTRICAS DE SIMILARIDADE DE PERFIL DE PESQUISADORES

Segundo o dicionário Oxford[15] (*Oxford Dictionaries*), métrica é um método para medir algo ou os resultados obtidos com este. É o conjunto de regras que orienta a medida.

[15] Disponível em http://www.oxforddictionaries.com.

Uma métrica determina padrões de medição que um indicador pode ser avaliado. É a própria medida (métrica direta) ou um método de cálculo (métrica indireta). As métricas científicas têm um papel importante na comunidade acadêmica, pois auxiliam no processo de medição da qualidade da produção científica e na identificação de especialistas em determinadas áreas. Tais métricas baseiam-se fortemente nas citações de artigos de pesquisadores. No entanto, para Cervi, Galante e Oliveira [5], [11], falta considerar a trajetória do pesquisador e o cenário onde ele está inserido. As citações a artigos deveriam ser um dos elementos a serem considerados no processo utilizando, assim, o perfil como um todo, dando pesos ou valores a determinados aspectos do currículo, favorecendo quesitos a serem avaliados. Isso daria uma maior adaptabilidade aos processos de análises, os quais se diferem um do outro.

Métricas como H-index[16] e G-index[17] calculam a importância do pesquisador de forma individual, baseando-se apenas em citações bibliográficas. Por exemplo, o H-index, é o número de artigos feito pelo pesquisador com citações maiores ou iguais a esse número. Todavia, este índice tem problemas para avaliar pesquisadores em início de carreira e, como toda tentativa simplista de se classificar pesquisador por um único número, o índice enfrenta várias críticas. Entre as críticas estão: diferenças entre áreas, diferenças entre idade, entre outros.

A métrica proposta por Cervi, Galante e Oliveira [5], [11], tem o objetivo de classificar o pesquisador em um nível de reputação definindo através de um índice. Esta normalização foi proposta com a finalidade de não distanciar com valores numéricos pesquisadores iniciantes, intermediários e experientes e tornar mais compreensível a identificação da reputação do pesquisador. Assim, a métrica é definida pelo somatório entre os elementos que compõem o modelo de perfil do pesquisador como também, são definidos pesos para as categorias apontadas no modelo. O somatório dos pesos dos elementos é limitado ao peso máximo de cada categoria e o próprio somatório tem um valor limitado.

Já os autores Zhang e Hurley [32], propuseram particionar o perfil do usuário em grupos de itens semelhantes e compor uma lista de recomendação de itens que correspondam, bem como cada *cluster* em que o perfil foi alocado. Isso, segundo os autores, aumentaria a possibilidade de semelhanças entre os itens utilizando métricas de similaridade como cosseno. Essa métrica é amplamente usada na área de recuperação de informação para calcular a similaridade entre vetores TF-IDF. O cálculo da similaridade entre usuários, é um artifício

[16] Medir o seu Impacto: Fator de Impacto, análise de citações e outras métricas. Disponível em http://researchguides.uic.edu/c.php?g=252299&p=1683205. Acessado em 07/11/2017.
[17] Medindo o impacto de sua pesquisa: G-Index. Disponível em http://guides.library.cornell.edu/c.php ?g=32272&p=203392. Acessado em 07/11/2017.

heurístico empregado para definir níveis de semelhança entre os perfis dos usuários, permitindo encontrar o grupo de usuários mais próximos chamado de vizinhança do usuário, os perfis semelhantes. O cálculo comporta a visualização de quão similar é as avaliações dos usuários baseado nas avaliações de itens.

Outra função de similaridade amplamente utilizada é Pearson, onde a medida do coeficiente de Pearson calcula a correlação entre os usuários. Esta medida foi pensada para resolver um problema que ocorre junto à similaridade baseada no cosseno, a qual não considera a diferença entre as notas que um usuário atribui a determinado item. Ou seja, um usuário pode ser mais rigoroso em suas avaliações dos itens do que outros.

2.5. SISTEMAS DE RECOMENDAÇÃO DE PESQUISADORES

Para Ricci, Rokach e Shapira [33], sistemas de recomendação são uma combinação de várias técnicas computacionais que personalizam itens com base nos interesses do usuário, segundo o contexto no qual estão inseridos. Os sistemas de recomendação são uma das aplicações de aprendizagem de máquina e têm por objetivo sugerir itens a um usuário com base em seu perfil [34]. Assim, comparando as preferências de um usuário com um grupo de usuários é possível fazer recomendações relevantes a esse usuário. Itens com características similares aos que o usuário já demonstrou interesse no passado também podem gerar recomendações relevantes no futuro. Nesse contexto, o perfil do usuário pode ser obtido implicitamente ou explicitamente. Na forma implícita, informações são obtidas através de opções de seu histórico passado e até mesmo com sua localidade geográfica. Já na forma explícita, é possível apurar suas preferências, utilizando *feedbacks* do próprio usuário.

Para ajudar a gerar sugestões relevantes para os pesquisadores, sistemas de recomendação buscam alavancar os interesses ocultos nos perfis de publicação dos próprios pesquisadores. Embora uma das informações mais utilizadas seja a rede de citação do pesquisador, a qual vem demonstrando melhor desempenho para a recomendação, a rede é frequentemente escassa, o que torna difícil recomendação. Os dados das citações são uma rica fonte de informação, porém, estão sujeitos a certas limitações. Um exemplo é a utilização em estudos que dependem exclusivamente da rede de citações de trabalhos. Mesmo que estes sejam considerados de ponta, são marginalizados por não terem nenhum citações logo no início. Este é um tipo de "problema de arranque a frio" em sistemas de recomendação

acadêmicos, o qual é análogo ao mesmo problema em sistemas de recomendação em geral [3].

Nascimento *et al.* [22] desenvolveram um sistema de recomendação de trabalhos acadêmicos em que utilizaram o título dos trabalhos para a construção do perfil de usuário, bem como o resumo para gerar vetores de características de trabalhos (*TAG*) candidatos a recomendar.

No trabalho de Sugiyama e Kan [3], ao contrário dos sistemas de recomendação que utilizam as citações somente para a criação do perfil, os dados são utilizados de forma mais eficaz, através de modelagens de potenciais citações, ou seja, reforçando a rede de citações, recomendando através de filtragem colaborativa citações que ajudem no trabalho em desenvolvimento.

Já Lee, Lee e Kim [6] propõem um sistema de recomendação de artigos acadêmicos personalizado. Utilizando um rastreador para recuperar artigos na web, examinam a semelhança entre dois trabalhos com base na semelhança dos textos. Assim, detecta-se automaticamente os temas de preferência do pesquisador e recomenda-se os artigos relacionados baseado na similaridade das obras. No processo de aprendizagem das preferências, aplicou-se o método do K vizinhos mais próximos (KNN) e, para esta tarefa de recomendação, utilizou-se *clustering* e algoritmo de recomendação baseado em vizinhança.

Gollapalli, Mitra e Giles [7] abordaram um sistema de recomendação de pesquisadores com instâncias similares para o domínio acadêmico. Através de uma consulta pelo nome do pesquisador, o objetivo do sistema é a recomendação da lista de pesquisadores que têm experiência semelhante ao do pesquisador consultado em suas áreas de especialização. Nesse caso, são propostos modelos de semelhança entre pesquisadores com base em perfis de especialização extraídos de suas publicações e suas páginas acadêmicas pessoais. Contrastando assim, trabalhos que utilizam a previsão de coautor para a recomendação, o objetivo é recuperar os pesquisadores que trabalham em áreas similares, mesmo se eles estão muito distantes da rede de coautoria. Dado um conjunto de perfis de especialização para os pesquisadores, foram exploradas as seguintes técnicas para a semelhança entre dois perfis:

- *Okapi BM25* (OKAPI), utilizada para o cálculo da similaridade entre dois perfis, tratando um perfil como a consulta e o segundo como um documento;
- *KL Divergence* (KLD), empregada para quantificar à similaridade entre duas distribuições de probabilidade;

- *Probabilistic Modeling* (PM), a similaridade entre dois perfis é calculada usando probabilidade condicional;
- Trace-based Similarity (REL), é calculado através de uma matriz de densidade relevante entre dois perfis.

Os resultados apontam a necessidade de técnicas para uma representação mais precisa dos perfis de pesquisador, não apenas em termos dos documentos, mas incluindo também a rede acadêmica subjacente entre pesquisadores e informações de metadados.

Em sistemas de recomendação, método de recomendação é a forma como as informações coletadas serão utilizadas para gerar recomendações. Atualmente, o método de recomendação é o tema que mais desperta interesse nos pesquisadores. Nesse sentido, as formas de recomendação existentes são seleção manual, resumo estatístico, baseada em atributos, correlação item-para-item e correlação usuário-para-usuário. Aqui destacam-se duas:

- Recomendação baseado em atributos, onde utiliza-se características do perfil do pesquisador para gerar recomendação. Por exemplo, uma pesquisa do pesquisador P1 poder ser recomendada a um pesquisador P2 que frequentemente publica artigos relacionados ao tema da pesquisa;
- Recomendação de correlação usuário-para-usuário, onde são recomendados para o usuário de acordo com a similaridade entre seu perfil e o perfil dos outros usuários. Se diversas pessoas com os mesmos hábitos gostaram de um mesmo assunto, é provável que o usuário a ser recomendado também goste.

Algumas técnicas predominam em sistemas de recomendação, como a filtragem colaborativa, filtragem baseada em conteúdo e sistemas híbridos. A filtragem colaborativa faz recomendações baseadas nas avaliações dos itens realizadas por um grupo de usuários “vizinhos”, cujos perfis de avaliações são os mais similares ao do usuário alvo. Por sua vez, a filtragem baseada em conteúdo recomenda itens ao usuário cujo conteúdo é similar ao conteúdo que o usuário tenha visto ou selecionado recentemente. Cada uma das técnicas possui vantagens e desvantagens dependendo do contexto e, assim, os sistemas híbridos visam utilizar um conjunto de duas ou mais técnicas de recomendação.

2.6. REDES SOCIAIS ACADÊMICAS

Uma rede social pode ser imaginada como grafo, onde G = (V, E). O G é o grafo formado por vértices (V) e arestas (E). Cada vértice representa um perfil e cada aresta representa a relação existente entre dois perfis integrantes da rede. A modelagem em rede vem sendo aplicada em diversas áreas como colaboração científica [35]. Dados modelados em rede podem ser analisados através de métricas de análise de rede social. Para Wasserman e Faust [36], a área de análise de rede social tem atraído muito interesse nas últimas décadas. Com as métricas de análise de redes sociais é possível identificar alguns aspectos, dentre eles; padrões de relacionamento entre os perfis de uma rede, a conectividade entre os mesmos, a formação de *clusters,* a evolução da rede ao longo do tempo e o fluxo de comunicação, informação e conhecimento dentro da rede.

Ainda, segundo Wasserman e Faust [36], os perfis mais importantes ou mais proeminentes estão normalmente localizados em posições estratégicas dentro da rede. A centralidade do grau de um perfil corresponde ao número de arestas incidentes ou ao número de vértices adjacentes a ele. Dessa forma, reflete-se a posição e o papel do perfil em termos de popularidade e atividade. Se as redes forem valoradas, onde a aresta possui um peso, pode-se levar em conta este valor, nas redes de coautoria essa medida determina o grau de colaboração de um ator.

Redes sociais acadêmicas destacam se no contexto de redes sociais de colaboração científica. Estas redes revelam como os pesquisadores se relacionam por meio do desenvolvimento de seus trabalhos e publicações. Para isso, utiliza-se de técnicas de mineração de dados, onde cria-se redes iniciais automaticamente através de informações disponíveis na web, como por exemplo, em páginas pessoais dos pesquisadores e em diferentes bibliotecas digitais. A interação do usuário com o mundo real produz informações para a obtenção de modelos de redes sociais e, assim, basta a descoberta e captura dessas interações dos usuários em sistemas disponíveis na web para montar as redes. Redes de colaboração científica entre pesquisadores são apresentadas e debatidas em trabalhos desenvolvidos como em Tang *et al.* [37]–[39], onde utilizam dados da produção científica de pesquisadores no contexto de redes sociais acadêmicas, considerando as relações que pesquisadores tem entre si. Basicamente, os trabalhos analisam redes de coautoria por meio da definição de grafos e suas relações.

No trabalho de Brandão e Moro [40], recomenda-se colaborações em redes sociais acadêmicas, baseando-se na afiliação dos pesquisadores. Em suas avaliações experimentais,

mediu-se a acurácia das recomendações, como também a novidade e a diversidade das recomendações. Definiu-se uma nova metodologia chamada de *Affin* e uma função de recomendação que combina a métrica de afiliação institucional e proximidade social. Os resultados mostram que as recomendações apresentam maior acurácia, novidade e diversidade quando se leva em consideração aspecto de afiliação institucional. Já no trabalho de Brandão, Moro e Almeida [41], é feita uma análise de fatores que impactam nas recomendações de colaboração acadêmica. As recomendações foram avaliadas quanto as métricas de revocação, novidade, diversidade e cobertura. Em sistemas de recomendação pode avaliar-se as recomendações por mecanismos de retorno do usuário. Porém, os autores avaliam que, no contexto acadêmico, tal retorno é complexo. Isso porque um pesquisador pode avaliar uma recomendação como ruim por questões subjetivas, como por exemplo, não gostar do recomendado, por falta de afinidade ou por competição. Desta forma, em seu trabalho, avaliaram as recomendações de colaboração dividindo a rede social em duas partes. Na primeira parte dos dados, aplicaram as funções de recomendação gerando a lista de recomendação. Na segunda metade dos dados, foi feito o comparativo entre as colaborações efetivamente realizadas. Como resultado, foi verificado se as recomendações feitas para a parte inicial dos dados estavam presentes na segunda metade dos dados, denominada de conjunto verdade.

2.7. REDES DE COAUTORIA ACADÊMICA

Uma rede de coautoria acadêmica, segundo Maia e Caregnato [42], mostra atividades acadêmicas na forma de produção bibliográfica, realizadas de forma conjunta por um grupo de pesquisadores. Em redes de coautorias acadêmicas os pesquisadores são considerados atores/entidades e as colaborações são consideradas relações/ligações entre os pesquisadores. Assim, podemos representar essa rede na forma de um grafo, onde o pesquisador representa um nó e as arestas entre dois nós representa pelo mesmo uma produção feita em coautoria.

É com esse enfoque que Mena-Chalco, Digiampietri e Cesar [8] empregam duas métricas individuais (características estimadas para cada pesquisador), referentes ao número de colaboradores e a medida que estima a colaboração no grupo. As produções bibliográficas para levar-se em consideração na identificação de redes de coautoria são os artigos publicados em periódico, livros publicados, capítulos de livros, textos em jornais de notícias ou revistas e publicações em anais de congressos. Já nas métricas globais, características estimadas sobre

todo o grafo, são consideradas doze métricas sobre a rede de coautoria. Essas métricas exploradas ajudam a elucidar a compreensão da estrutura e dinâmica de colaboração em redes de coautoria acadêmica. Aqui destacam-se duas:

- **Transitividade**: Refere-se à probabilidade de que dois nós adjacentes a um nó estejam ligados. Usualmente, este valor também é chamado de coeficiente de *clustering*, e para o nó *i* este valor representa a proporção de arestas entre nós dentro da vizinhança do nó *i*, dividido pelo número máximo de arestas que poderiam existir entre todos os vizinhos. Isto é, o coeficiente de *clustering* é a razão entre o número de triângulos que contêm o nó *i* e o número de triângulos que poderia existir se todos os vizinhos do nó *i* forem interligados.
- **Assortatividade**: Para o nó *i* refere-se à preferência do nó *i* ter arestas a outros nós que mantenham grau do nó similares (ou diferentes) ao grau do nó *i*. O coeficiente de assortatividade é o coeficiente de correlação de Pearson dos graus entre os pares de nós ligados. Valores positivos do coeficiente de Pearson indicam uma correlação entre nós de grau similar, entretanto valores negativos indicam uma correlação entre nós de grau diferente. Este coeficiente pode variar de um mínimo de -1 a um máximo de 1 (rede com assortatividade máxima).

Uma rede de coautoria entre pesquisadores, também chamada de rede de colaboração científica, é um exemplo reservado de uma rede social. A análise de redes sociais baseia-se na premissa que as relações entre os atores sociais podem ser descritas mediante um grafo, direcionado ou não-direcionado. Neste contexto, a vantagem da representação das redes de coautoria através de grafos permite a utilização da Teoria dos Grafos. Assim, pode ser feito uma análise dos comportamentos de relacionamento social, padrões e implicações destes relacionamentos, interpretando de forma mais qualificada suas relações dentro da rede [36].

É com esses princípios que Jr, Laender e Moro [43] modelam uma rede de coautoria como um grafo não direcionado, ponderado, no qual cada nó representa um autor e uma aresta entre dois nós representa que os autores publicaram artigo em coautoria. O peso inferido nas arestas é relativo ao número de artigos publicados em coautoria e representa a intensidade da colaboração entre os autores. Neste trabalho, utilizaram apenas os artigos escritos em conjunto para a coautoria, porém qualquer produção pode ser considerada como uma forma da documentação da colaboração entre dois ou mais autores.

No trabalho de Barbosa *et al.* [44], propõe-se uma ferramenta web para visualização e recomendação em redes de coautoria. A ferramenta possibilita fornecendo os dados das publicações, uma análise e recomendações visuais sobre a rede construída automaticamente com os dados da entrada e que também possibilita, ainda, uma análise aprofundada de uma rede de coautoria. Outra ferramenta online é apresentada no trabalho de Diniz *et al.* [45], a qual exibe recomendações de colaboração para pesquisadores, utilizando métricas de redes sociais.

2.8. TRABALHOS RELACIONADOS

Os trabalhos analisados possuem relação com os temas similaridade de perfil de pesquisadores e recomendação acadêmica. Desta forma, para uma melhor visualização, é feita uma comparação demonstrando os pontos em aberto de cada um deles através de critérios elencados e dispostos em tabelas.

2.8.1. Similaridade de Perfil de Pesquisadores

No trabalho de Lima *et al.* [2], foi proposta uma avaliação do desempenho dos principais pesquisadores em ciência da computação no Brasil, considerando os dados de sua carreira e não somente publicações e citações. Assim, utilizam tanto os dados da plataforma Lattes quanto da DBLP. Os resultados demonstraram a necessidade de considerar as peculiaridades dos perfis e não somente dados de suas publicações. Não foram utilizados aspectos de semelhança para determinar comparativos entre os perfis.

Em Wainer e Vieira [19] são analisadas decisões sobre os pesquisadores brasileiros e computadas suas correlações com 21 medidas diferentes. Ressalta-se que as métricas para avaliação da produtividade são baseadas em produção, produtividade e impacto do trabalho do pesquisador e não em similaridade entre os perfis dos pesquisadores.

No trabalho Vivian e Cervi [21], o perfil do pesquisador é construído com os dados da plataforma Lattes, de citações e da rede de coautoria, identificando o perfil através de técnicas de *Data Science*. Os autores utilizam uma métrica com pesos para melhor adaptabilidade do perfil permitindo aumentar a qualidade da classificação dos pesquisadores.

Já Mena-Chalco, Digiampietri e Cesar [8] utilizam-se dos dados do Lattes para a caracterização de uma rede de coautoria através do perfil do pesquisador. Uma comparação é feita em suas produções para gerar o grafo de coautoria. A distância de Levenshtein é

utilizada somente para obter a similaridade entre o nome e os títulos das produções devido a inconsistências nos dados, não utilizando de similaridade para verificação do perfil e sim para os dados. Da mesma forma, Hannel *et al.* [4] utilizam como função de similaridade o algoritmo de Smith-Waterman, utilizando após as extração das informações para comparar possíveis erros entre os títulos dos trabalhos.

Em Cervi, Galante e Oliveira [1], utilizam um conjunto de dados coletados de diferentes fontes como Plataforma Lattes, DBLP, Microsoft Academic Search e ArnetMiner para compor o perfil dos pesquisadores. Utilizou-se um modelo de perfil de pesquisador denominado Rep-Model que visa especificar o perfil do pesquisador de forma abrangente, envolvendo diversos elementos da carreira cientifica do pesquisador. Também foi utilizado uma métrica denominada Rep-Index, empregada para a classificação dos pesquisadores por sua reputação. A diferença da métrica Rep-Index em relação a outras é sua abrangência e adaptabilidade. A abrangência envolve a avaliação da reputação de pesquisadores, ponderando toda a trajetória científica estabelecida ao longo da carreira. Enquanto a adaptabilidade permite ao usuário utilizar uma abordagem em diferentes áreas e contextos, adaptando o modelo de perfil de acordo com critérios de sua necessidade. Essa abordagem torna-se flexível para as características das diversas área. Deste modo, pode ser utilizada mudando os pesos dos elementos que atenda as premissas da área e do contexto de utilização. Tal abordagem permite também avaliações por categorias, avaliando a reputação de pesquisadores levando em consideração somente as categorias desejadas.

Em Gollapalli, Mitra e Giles [7] são propostos modelos de semelhança entre pesquisadores com base em perfis de especialização extraídos de suas publicações e suas páginas acadêmicas pessoais. Dado um conjunto de perfis dos pesquisadores, foram exploradas as seguintes técnicas para a semelhança:

- OKAPI BM25: O perfil do pesquisador é representado utilizando um vetor correspondente a termos de um vocabulário determinado com base no conteúdo associado com o pesquisador. A similaridade entre os dois perfis é calculada usando a função BM25. A função é utilizada em recuperação de informação que classifica um conjunto de documentos, tratando um perfil como a consulta e o segundo como um documento;
- KLD (Kullback-Leibler Divergence): O perfil do pesquisador é representado como termos de uma distribuição de probabilidade. Por exemplo, dado um conjunto de documentos associados a um pesquisador, uma distribuição multinomial pode ser aplicada através da divergência de

Kullback-Leibler, utilizado para quantificar à similaridade entre duas distribuições de probabilidade;

- PM (*Probabilistic Modeling*): Pesquisadores tendem a trabalhar em áreas relacionadas e que poderiam ser mais apropriadas para modelar seus perfis como uma mistura de tópicos em vez de uma única distribuição multinomial. Alocação latente de Dirichlet (LDA - Latent Dirichlet Allocation) é uma ferramenta de modelagem de tema comumente utilizada para o agrupamento não supervisionado de dados e análise exploratória;
- REL (Trace-based Similarity): Calcula-se as matrizes densidade como um vetor unitário TF-IDF unidimensional representando o perfil de um pesquisador.

Mesmo utilizando técnicas diferentes das habituais, como o cosseno para medir a similaridade, o trabalho não utilizou-se de características mais profundas do perfil do pesquisadores, tratando-se muito do que a maioria dos trabalhos já utiliza, ou seja, dados das publicações e poucos dados encontrados em suas páginas pessoais. Nas conclusões do trabalho, observou-se a necessidade de explorar mais dados para uma melhor representação do perfil do pesquisador através de técnicas mais precisas. Também observa-se que não utilizaram ponderação para a medição da similaridade entre os perfis.

2.8.2. Recomendação Acadêmica

Na pesquisa proposta por Sugiyama e Kan [3], o perfil do pesquisador é constituído utilizando as listas de publicações da plataforma DBLP. Foi utilizada a abordagem de filtragem colaborativa para encontrar potenciais artigos a serem recomendados. Já Nascimento *et al.* [22] desenvolveram um sistema de recomendação de trabalhos acadêmicos, em que utilizaram o título e o resumo dos trabalhos para a construção do perfil de usuário, gerando vetores de características de trabalhos candidatos a recomendar. Por sua vez, Hong *et al.* [9] propõem um sistema de recomendação de artigos onde é extraído da palavras-chave dos artigos pesquisados pelo pesquisador, para compor um perfil que é atualizado a cada busca do usuário como também é calculado à similaridade do cosseno entre determinado tema e os artigos recolhidos.

Já Lee, Lee e Kim [6], propõem um sistema de recomendação de artigos acadêmicos utilizando um rastreador para recuperar artigos na web, examinando os trabalhos com base na semelhança dos textos e recomenda os artigos através do processo de K vizinhos

mais próximos (KNN). Não é utilizado nenhum outro atributo do perfil do pesquisador para a recomendação, nem é feito nenhuma ponderação junto a medida de similaridade.

No trabalho de Gollapalli, Mitra e Giles [7], é abordado um sistema de recomendação de pesquisadores através de uma consulta pelo nome do pesquisador. O objetivo do sistema é a recomendação da lista de pesquisadores que têm experiência semelhante ao do pesquisador consultado. Deste trabalho, ressalta-se o contraste com outros trabalhos, os quais no domínio acadêmico de recomendação acabam sugerindo artigos científicos para os usuário. Contrasta, também, com trabalhos que visam correlação junto a rede de coautoria dos pesquisadores, neste visa-se recomendar pesquisadores que trabalham em áreas similares mesmo que distante em suas redes de coautoria.

2.8.3. Comparação dos Trabalhos Relacionados

Com o objetivo de possibilitar uma visualização mais qualificada dos trabalhos relacionados que envolvem os campos de similaridade de perfil de pesquisadores e recomendação acadêmica, foram elencados alguns critérios de comparação dos trabalhos e dispostos em tabelas.

A Tabela 1 apresenta uma classificação dos trabalhos analisados com relação ao domínio dos dados para modelar o perfil do pesquisador. Foram definidos critérios de comparação julgados relevantes para o tema pesquisado conforme descrito a seguir:

- **Base Utilizada**: Quais base de dados de produção científica foram utilizadas para a coleta de dados, como por exemplo Plataforma Lattes, DBLP, Google Acadêmico e Microsoft Academic Search;
- **Tipo de Detecção**: Forma de detecção do perfil, podendo ser abordagem explícita, implícita ou híbrida. A detecção explícita exige intervenção do usuário, enquanto a implícita o próprio sistema define o perfil sem a participação direta do usuário. A híbrida é a união das duas abordagens;
- **Atributos do Perfil**: Quais atributos definiram o perfil do pesquisador, referindo-se a trabalhos publicados, currículo, citações, dentre outros;
- **Forma de Representação**: Apresenta a forma de representação do perfil pelo viés da modelagem de dados, sendo considerados vetores de características (*Tags*), ontologias, arquivos XML, dentre outros.

Tabela 1. Comparação entre os trabalhos relacionados ao domínio dos dados do perfil de pesquisadores.

Ref.	Base Utilizada	Tipo Detecção	Atributos do Perfil	Forma de Representação
[2]	Lattes e DBLP	Implícita	Dados do currículo, trabalhos publicados e citações	Conjunto de dados
[19]	Lattes, Web of Science, Scopus e Google Acadêmico	Implícita	Dados do currículo, trabalhos publicados e citações	Conjunto de dados
[21]	Lattes.	Implícita	Dados do currículo	Arquivo XML
[3]	DBLP	Implícita	Citações e fragmentos dos trabalhos como palavras-chave, resumo, introdução e conclusão	Vetor de termos
[22]	ACM, IEEE e ScienceDirect	Implícita	Título e o resumo dos trabalhos publicados	Vetor de termos
[8]	Lattes	Implícita	Dados do currículo	Conjunto de dados
[4]	Lattes, Critérios do CNPQ e ACM	Implícita	Dados do currículo	Ontologia
[9]	Google Acadêmico	Híbrida	Palavras-chave e texto do artigo	Arquivo XML
[6]	IEEE Xplore e ACM Digital Library	Implícita	Título, palavras-chave e o resumo dos trabalhos publicados	Vetor de termos
[7]	CiteSeerX	Implícita	Trabalhos publicados e Homepages Acadêmicas	Vetor de termos
[1]	Lattes, DBLP, Microsoft Academic Search, ArnetMiner	Implícita	Dados do currículo, trabalhos publicados e citações	Vetor de termos

Conforme a Tabela 1, é possível observar que a maioria dos trabalhos representam o perfil do pesquisador através de um vetor de termos, as quais são retiradas das mais diversas bases de dados sobre produção científica. É uma constante a utilização dos textos do trabalho do pesquisador e de citações para compor os atributos do perfil, dotando-se de técnicas para a identificação das principais palavras que representem melhor o texto e o trabalho do pesquisador. Porém, esses dados acabam produzindo um perfil pouco representativo do contexto total da vida do pesquisador impactando, assim, na tarefa de encontrar similaridades entre os perfis.

A Tabela 2 apresenta um estudo comparativo das características consideradas importantes quando se trabalha com similaridade de perfil. Foram definidos critérios de comparação julgados relevantes para o tema pesquisado conforme descrito a seguir:

- **Métrica de Similaridade**: Se a abordagem utiliza alguma métrica de similaridade de perfil ou de algum atributo do perfil, como cálculo de similaridade do cosseno, distância de Levenshtein, distância Jaro-Winkler, dentre outros;
- **Ponderação:** Se o trabalho incorpora alguma técnica de pesos em relação à similaridade do perfil;
- **Recomendação**: Apresenta se a abordagem utiliza algum mecanismo de recomendação e qual objeto é recomendado, como um evento científico, um outro pesquisador ou um artigo científico.

Tabela 2. Comparação entre os trabalhos relacionados ao domínio de similaridade de perfil e recomendação de pesquisadores.

Ref.	Métrica de Similaridade	Ponderação	Recomendação
[2]	Não	Não	Não
[19]	Não	Não	Não
[21]	Não	Não	Não
[3]	Não	Não	Artigo Científico
[22]	Não	Não	Artigo Científico
[8]	Distância de Levenshtein	Não	Não
[4]	Algoritmo de Smith-Waterman	Não	Não
[9]	Cosseno	Não	Artigo Científico
[6]	Cosseno	Não	Artigo Científico
[7]	OKAPI BM25, KLD, PM, REL	Não	Lista de pesquisadores
[1]	Não	Não	Não

Conforme a Tabela 2, observa-se que a similaridade de perfil de pesquisadores é utilizada para sugerir trabalhos relacionado ao do perfil alvo. Isso ajuda ao pesquisador a manter-se atualizado com um menor esforço na busca de trabalhos relacionados. Porém, outros elementos da carreira do pesquisador não são considerados para recomendações, o que também ajudariam nesse aspecto, como por exemplo, um evento científico ou um outro pesquisador. As medidas de similaridades são utilizadas de forma específica em determinados

aspectos de um atributo medido e não do perfil como um todo. Em nenhum momento, contudo, há uma ponderação entre quais aspectos representariam melhor a similaridade entre os perfis dos pesquisadores.

2.9. CONSIDERAÇÕES FINAIS DO CAPÍTULO

Como já ressaltado, este capítulo teve por objetivo apresentar uma visão geral sobre os conceitos essenciais ao tema similaridade de perfil de pesquisador e recomendação acadêmica, bem como explanar as principais características do perfil de pesquisador e exemplificar algumas bases de dados científicas.

Inicialmente, foi apresentado que estudos sobre a produção científica de pesquisadores estão ganhando espaço, muito pela necessidade de qualificação e análise da produção científica para a tomada de decisões de fomento. Assim, as bases de produções científicas estão ganhando visibilidade para a verificação do perfil do pesquisador e várias iniciativas para manter de forma organizada os dados produzidos pelas pesquisas. Ao mesmo tempo surgiram os problemas como a falta de filtros dentro das plataformas para a captura do perfil dos pesquisadores e análise de como as pesquisas podem se relacionar. Neste contexto, cresce a necessidade de ferramentas para manipular os dados e avaliar os perfis, com o objetivo de analisar, orientar e sugerir novas conexões entre pesquisas, pesquisadores e agências de fomento.

Na sequência, tratou-se de compreender o que é um perfil de usuário e como modelar sua representação, através de alguns serviços e trabalhos que trataram sobre perfil. Em seguida, apresentou-se o conceito de similaridade aplicada a perfil, como se estabelece e como está sendo utilizada em alguns trabalhos desenvolvidos. Também foi estudado o conceito de métrica, para entender como as métricas de similaridade estão sendo utilizada nos trabalhos atuais ligados ao perfil de pesquisador.

Procurou-se, ainda, elucidar conceitos sobre sistemas de recomendação no contexto acadêmico, verificando-se as diversas técnicas para personalizar itens com base nos interesses dos usuários. Foi levantado quais dados utilizam e o que recomendam para os pesquisadores. Redes sociais científicas foram estudadas e foram verificadas como as relações entre pesquisadores influenciam em seu trabalho. Com relação ao contexto de uma rede de coautoria acadêmica, foi identificado como a mesma é formada pela relação das citações dos trabalhos e suas informações das produções bibliográficas dos diversos pesquisadores.

Os trabalhos analisados demonstraram que existem muitos desafios a serem explorados no tema que envolve similaridade de perfil de pesquisador e recomendação acadêmica. Nesse sentido, é fato que as pesquisas científicas sobre o tema tendem a aumentar, tendo em vista que a humanidade está cada vez mais produzindo informação e a quantidade a ser avaliada e explorada tende a dificultar a experiência dos pesquisadores e sua atualização frente a seus temas de pesquisa. Diante disso, cria-se um espaço importante para que soluções tecnológicas sejam implementadas, em diversas áreas de pesquisa, como aplicações para dispositivos móveis, adaptabilidade frente ao contexto, recomendação de informações desejadas pelo usuário.

Muitos dos trabalhos utilizaram à similaridade para a desambiguação dos dados bibliográficos ou para a verificação da semelhança entre *strings* retiradas dos textos dos trabalhos dos pesquisadores, especialmente para verificar o quão semelhantes são e, assim, sugerir estas a outros pesquisadores. As sugestões parecem ser bastante centradas em relação aos trabalhos bibliográficos e não em outros aspectos que compõem a totalidade do perfil do pesquisador. Também, não se trabalha aspectos relacionados a ponderação dos atributos referentes à similaridade do perfil do pesquisador.

Após a análise dos resultados, constatou-se que existem diferentes técnicas para medir a similaridade dos perfis. No entanto, por mais que estes métodos tenham sidos bem sucedidos em casos específicos onde foram utilizados, não foi possível detectar uma forma padronizada para avaliar a similaridade de um perfil de pesquisador. Dentro deste contexto, a extração de característica e medida de similaridade é um desafio, devido a uma grande variedade de funções de medida de similaridade, que podem ser combinadas com as diferentes técnicas que retornam resultados que nem sempre são os mais satisfatórios.

Destaca-se que é necessário detectar quais variáveis são relevantes quanto a avaliação de similaridade em perfil de pesquisadores, ou seja, quais os principais fatores, características ou atributos são necessários e como medi-los, para obter com mais precisão um índice de similaridade do perfil de um pesquisador, gerando assim uma recomendação com mais qualidade.

3. ABORDAGEM PROPOSTA

O presente capítulo tem por objetivo apresentar a abordagem proposta, destacando suas características. Assim, na primeira Seção, indicada como 3.1, é apresentado uma visão macro do que se pretende. Já na Seção indicada como 3.2, apresenta-se o modelo de perfil de pesquisadores e descreve-se suas adaptações para avaliação das similaridades dos perfis. Na sequência do trabalho, na Seção 3.3, descreve-se o modelo de similaridade de perfil de pesquisadores e seus índices apresentados por meio de equações.

As regras e funções de recomendação são apresentadas na Seção 3.4. Tal sessão é seguida pela Seção 3.5, que descreve o processo de delimitação dos dados utilizados, como foram coletados e preparados para uso. Por fim, na Seção 3.6, são apresentados o desenvolvimento de uma aplicação web integrando o modelo de perfil, os cálculos de similaridade e as métricas de recomendação.

3.1. VISÃO GERAL

A abordagem desenvolvida permite a identificação de perfis de pesquisadores similares, baseando-se em dados gerais dos currículos e não somente em produções científicas. A abordagem oportuniza, além da busca por perfis similares, a recomendação de perfis e trabalhos científicos realizados por perfis próximos ao do usuário, sendo abrangente e independente do domínio. A abrangência da abordagem se dá pela quantidade e diversidade de elementos da carreira do pesquisador. Já a independência de domínio possibilita que a abordagem possa ser utilizada em qualquer contexto ou área do conhecimento.

Neste estudo, os dados do usuário foram importados diretamente da Plataforma Lattes, uma vez que é uma base de dados rica em informações de pesquisadores, especialmente porque é o próprio pesquisador que faz a inserção dos dados, o que pode minimizar a ocorrência de erros. Cabe destacar que a abordagem proposta possibilita que outras bases de dados possam ser utilizadas, bastando, para isso, que contemplem o modelo de perfil a ser utilizado.

A identificação dos perfis similares de pesquisadores é uma das contribuições que se objetiva a partir da produção desse trabalho, através de uma abordagem para comparar perfis de pesquisadores, em conjunto com a proposição de uma métrica para calcular a similaridade entre eles. Também foi desenvolvida uma aplicação web com a métrica e o

mecanismo de recomendação, conforme apresenta a Figura 8, que mostra a arquitetura da solução. Nessa mesma figura, apresenta-se a visão geral do trabalho, mostrando o fluxo único de uma carga de dados para iniciar a aplicação, tentando assim melhorar o problema de arranque frio que se observa em sistemas de recomendação, da mesma forma que deixa calculado os dados de similaridade de um grande número de registros. Desta forma, facilita-se as atualizações dos currículos posteriores coletados, embora deva-se ressaltar que a cada currículo atualizado, um novo cálculo é gerado com todas as comparações que existem na base de dados em relação a este currículo.

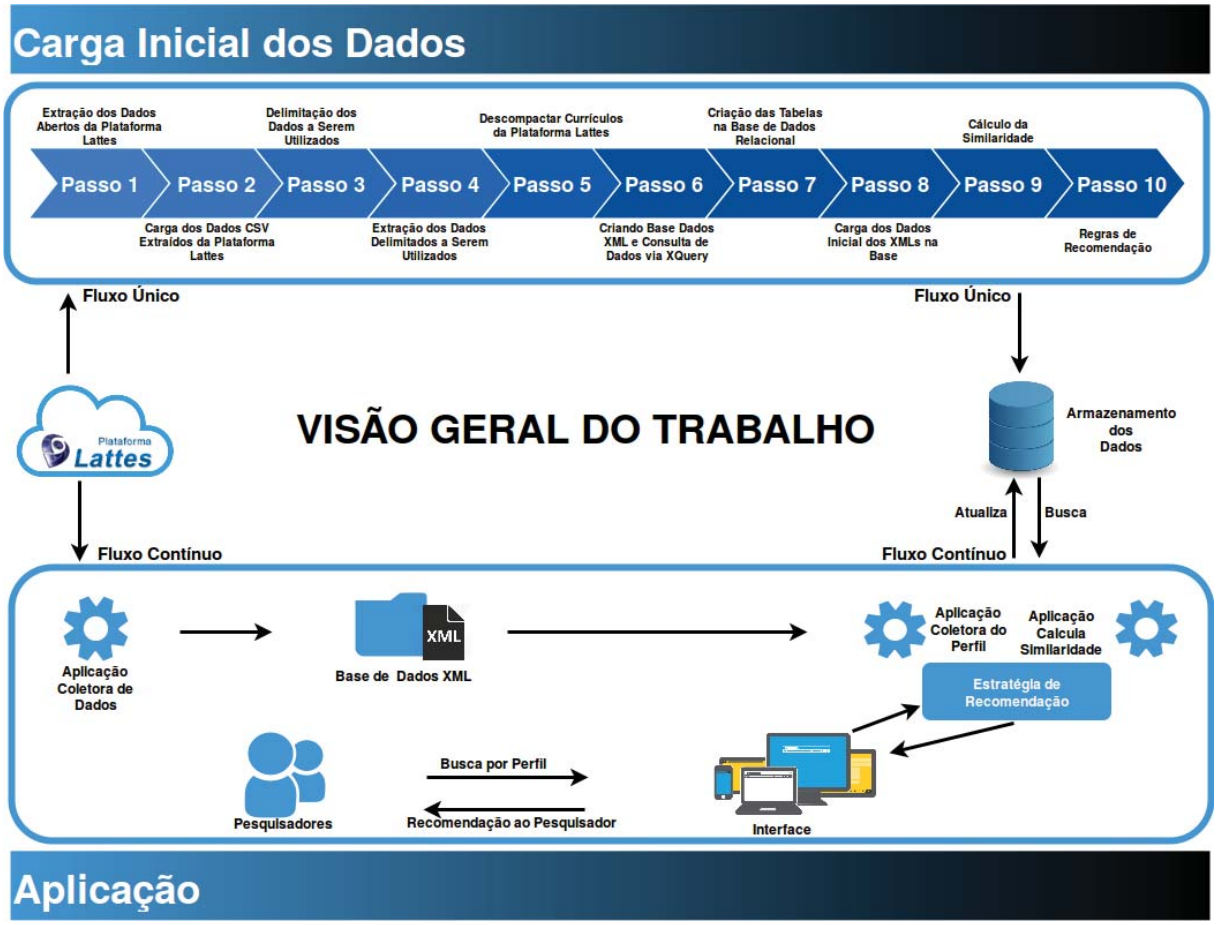


Figura 8: Visão geral da aplicação web, com a carga inicial dos dados, o cálculo da métrica e o mecanismo de recomendação.

É imprescindível destacar que esta é apenas uma visão geral da abordagem proposta, já que todo o processo da carga inicial dos dados e a aplicação serão melhor ilustrados na sequência da dissertação.

3.2. MODELO DE PERFIL DE PESQUISADORES

Para o modelo de perfil de pesquisadores utilizou-se como base o modelo proposto por Cervi, Galante e Oliveira em [1], [5], [11] denominado Rep-Model. O Rep-Model elenca diversos elementos para a identificação do perfil do pesquisador, sendo que alguns desses elementos são amplamente utilizados nas comunidades científicas das diversas áreas do conhecimento. A abordagem do modelo, com suas categorias, seus elementos e siglas são apresentadas na Tabela 3.

Tabela 3. Categorias, elementos e siglas da abordagem Rep-Model [5].

Category	**Elements**	**Acronym**
Identification (ID)	Name	NM
	Institution	INST
	Education Degree	ED
Advisory (ADV)	Master Dissertation Advisor	MDA
	PhD Thesis Advisor	PTA
	Postdoctoral Advisor	PA
Examining Board (EB)	Participation in Examination Boards Master Dissertation	PEBMD
	Participation in Examination Boards PhD Thesis	PEBPT
Membership (MS)	Conference Committee Coordinator	CCC
	Conference Committee Member	CCM
	Editorial Board Member	EBM
	Reviewer of Journals	RJ
Production (PROD)	Articles in Scientific Journals	ASJ
	Book Chapter Published	BCP
	Book Published	BP
	Complete Work Published in Conference Proceedings	CWPCP
	H-Index	HI
	Network Co-authorship	NC
	Research Projects	RP
	Software	SOFT

Para melhor compreensão dos dados envolvidos na modelagem do perfil do pesquisador, descreve-se, a seguir, o modelo apresentado na Tabela 3, especificando melhor cada elemento das categorias [5].

Identification (ID): categoria que identifica o pesquisador através dos seguintes elementos: (i) Elemento *Name* representa o nome do pesquisador; (ii) Elemento *Institution* representa a instituição de vínculo do pesquisador; (iii) Elemento *Education Degree* representa a titulação do pesquisador, tendo como opção mestrado, doutorado ou pós-doutorado.

Advisory (ADV): categoria que representa as orientações de trabalhos do pesquisador da seguinte forma: (i) Elemento *Master Dissertation Advisor* representa as orientações de mestrado; (ii) Elemento *PhD Thesis Advisor* representa as orientações de doutorado; (iii) Elemento *Postdoctoral Advisor* representa as orientações de pós-doutorado.

Examining Board (EB): categoria que especifica a participação do pesquisador em bancas de trabalhos da seguinte forma: (i) Elemento *Participation in Examination Boards Master Dissertation* representa a participação em defesas de mestrado; (ii) Elemento *Participation in Examination Boards PhD Thesis* representa as participações em defesas de doutorado.

Membership (MS): categoria que representa as conferências onde o pesquisador foi coordenador ou membro de comitê de programa e também representa onde o pesquisador foi revisor de periódico ou membro de corpo editorial, da seguinte forma: (i) Elemento *Conference Committee Coordinator* representa as conferências onde o pesquisador foi coordenador de comitê de programa; (ii) Elemento *Conference Committee Member* representa as conferências onde o pesquisador teve participação como revisor de trabalhos; (iii) Elemento *Editorial Board Member* representa as atuações do pesquisador como membro de corpo editorial de periódicos; (iv) Elemento *Reviewer of Journals* representa as atuações do pesquisador em revisões de periódicos.

Production (PROD): representa as produções científicas e outros elementos relacionados a produção científica do pesquisador, da seguinte forma: (i) Elemento *Articles in Scientific Journals* representa os artigos publicados em periódicos; (ii) Elemento *Book Chapter Published* e (iii) Elemento *Books Published* representam, respectivamente, os capítulos de livros e os livros publicados pelo pesquisador; (iv) Elemento *Complete Work Published in Conference Proceedings* representa os trabalhos completos publicados em conferências; (v) Elemento *H-Index* representa às citações dos artigos do pesquisador por outros pesquisadores; (vi) Elemento *Network Co-authorship* representa o número de

pesquisadores que possuem trabalhos publicados em conjunto com o pesquisador em questão; (vii) Elemento *Research Projects* representa o número de projetos de pesquisa do pesquisador; (viii) Elemento *Software* representa o número de softwares em que o pesquisador teve participação.

O Rep-Model define diferentes pesos para as categorias e seus elementos. O somatório dos pesos totaliza o valor 100. Os pesos das categorias e os pesos dos elementos são definidos pelo utilizador, onde pode levar em consideração as particularidades e especificidades de cada área. Por essa característica, o modelo torna-se bastante adaptável ao contexto que se necessita.

Já a Tabela 4, apresenta um exemplo de pesos para as categorias e os elementos do modelo, da mesma forma que, indica o maior valor de cada elemento.

Tabela 4. Exemplo do Rep-Model com pesos e valores máximos de intervalo [11].

Category	**Weight**	**Elements**		**Weight**	**Maximum Value of Range**
Identification (ID)	15	Name		-	-
		Institution		-	-
		Education Degree	Postdoctoral	15	-
			Doctorate	12	-
			Master	8	-
Advisory (ADV)	15	Master Dissertation Advisor		4	66
		PhD Thesis Advisor		5	30
		Postdoctoral Advisor		6	4
Examining Board (EB)	15	Participation in Examination Boards Master Dissertation		5	72
		Participation in Examination Boards PhD Thesis		10	26
Membership (MS)	15	Conference Committee Coordinator		3	9
		Conference Committee Member		2	107
		Editorial Board Member		6	5
		Reviewer of Journals		4	32
Production (PROD)	40	Articles in Scientific Journals		12	96
		Book Chapter Published		3	35
		Book Published		5	15
		Complete Work Published in Conference Proceedings		8	437
		H-Index		6	24

Category	Weight	Elements	Weight	Maximum Value of Range
		Network Co-authorship	2	132
		Research Projects	1	37
		Software	3	21
Total	**100**	**Total**	**100**	

Observa-se no elemento *Education Degree*, que pertence a categoria *Identification*, uma estrutura decisória com opções, onde apenas uma deve ser utilizada (Pós-doutorado / Doutorado / Mestrado) tendo um peso máximo de 15, neste exemplo.

3.2.1. Adaptações no Modelo de Perfil de Pesquisadores

A escolha do Rep-Model como modelo de perfil de pesquisadores utilizado nesse trabalho, justifica-se por se tratar de um modelo de perfil abrangente e adaptável, muito embora tenham sido realizadas algumas adequações necessárias ao contexto de similaridade. Dentre essas adequações, foram acrescentados três novos elementos ao Rep-Model, a saber: (i) áreas do conhecimento do pesquisador; (ii) áreas de atuação do pesquisador; e (iii) linhas de pesquisas do pesquisador. Com base no entendimento pretendido nesta abordagem, não faz sentido comparar, por exemplo, a produção quantitativa de pesquisadores de diferentes áreas do conhecimento, já que cada área tem suas peculiaridades, e algumas desenvolvem mais artigos para periódicos do que outras. Da mesma forma, algumas áreas potencializam a produção técnica, como registro de software e patentes, e outras focam em publicações em conferências.

Diante desse contexto de diversidade de áreas e suas especificidades, também não é interessante uma similaridade somente pela produção quantitativa dos pesquisadores. Por estes motivos foram retirados dos elementos a parte qualitativa, aplicando a extração textual de seu conteúdo. Isto porque o modelo original do Rep-Model foi concebido para identificar a reputação do pesquisador, na qual os elementos do modelo são números inteiros de sua produção e vida acadêmica. Já para a similaridade, também busca-se saber de cada elemento o que foi produzido e não somente o quanto; bem como quais termos são relevantes, ou seja, as características que envolvem o trabalho do pesquisador.

O elemento H-Index, que representa às citações dos artigos do pesquisador por outros pesquisadores, não foi utilizado na abordagem proposta, visto que essa informação não

está contida nos dados dos XMLs dos pesquisadores coletados da Plataforma Lattes. Dessa forma, esse elemento poderá ser inserido em trabalho futuro, o qual poderá definir, por exemplo, de onde e como será possível buscar essa informação para a base de dados. Aliado a isso, o elemento H-Index não foi avaliado como essencial para identificar a similaridade entre pesquisadores, uma vez que é um valor quantitativo, sem representar área, contexto e especificidades.

A modelagem do perfil do pesquisador é baseada em conhecimento, usando dados coletados da Plataforma Lattes. O Rep-Model indica o que é importante buscar dentro do currículo do pesquisador, determinando pesos para isso. Busca-se então, as informações quantitativas dos elementos, exemplo, total de artigos produzidos. Bem como as informações qualitativas, exemplo, os termos dos artigos produzidos. Assim, temos a informação do quanto foi produzido e o que foi produzido. Dessa forma, de acordo com o comportamento ou interesse do pesquisador, demonstrado nos elementos informados no Lattes, é possível fazer inferências sobre suas preferências a partir da vida acadêmica. Pode-se citar como exemplo, se o pesquisador publica sobre banco de dados, possivelmente, será interessante para o mesmo receber recomendações sobre esta área. É com base nisto que a Tabela 5 mostra os elementos do perfil do pesquisador já com as adaptações necessárias.

Tabela 5. Modelo de perfil de pesquisador adaptado para similaridades.

Categoria	**Sigla**	**Elementos**		**Sigla**
Pesquisador	PES	Nome		NM
		Instituição		INST
		Áreas de Conhecimento		AC
		Áreas de Atuação		AA
		Linhas de Pesquisas		LP
		Formação Acadêmica	Pós-Doutorado	FA
			Doutorado	
			Mestrado	
Orientação	ORI	Orientação de Mestrado		OM
		Orientação de Doutorado		OD
		Orientação de Pós-Doutorado		OP
Banca	BAN	Participação em Banca de Mestrado		PBM
		Participação em Banca de Doutorado		PBD
Comitê	COM	Coordenação de Comitê de Conferência		CCC

Categoria	Sigla	Elementos	Sigla
		Membro de Comitê de Conferência	MCC
		Membro de Corpo Editorial de Periódico	MCEP
		Revisão de Periódico	RP
Produção	PRO	Artigos em Periódicos	AP
		Capítulos de Livros	CL
		Livros	LIV
		Trabalhos Completo em Conferências	TCC
		Rede de Coautoria	RC
		Projeto de Pesquisas	PP
		Softwares	SOFT

Os três elementos adicionados não recebem pesos, visto que são utilizados para identificação de áreas de trabalhos, na conjuntura de produção qualitativa, o que permite verificar a similaridade dentro deste contexto, contribuindo para a formação da similaridade através das palavras-chave que contém.

3.3. MODELO DE SIMILARIDADE DE PERFIL DE PESQUISADORES

O modelo de similaridade foi desenvolvido utilizando dados dos elementos do perfil do pesquisador, sendo um aspecto quantitativo, baseado na produção acadêmica e, outro qualitativo, fundamentado no que o pesquisador desenvolveu durante sua carreira. Tais elementos são utilizados para a verificação da similaridade de características, sendo a importância de cada uma delas atribuída usando ponderação. Nesse contexto, a principal função do modelo de similaridade é definir como as características são comparadas para a obtenção do grau de similaridade. Assim, a abordagem proposta trabalha com similaridade global, onde se observa todos os elementos do perfil, e a similaridade local, podendo ser observado um elemento específico ou uma categoria do modelo de perfil de pesquisadores.

Com um modelo abrangente, torna-se necessário similaridades parciais para compor um valor final que indicará a similaridade entre dois perfis. Podem ser destacados como os indicadores de similaridade: (i) a similaridade quantitativa, baseada na produção do pesquisador; e (ii) a similaridade qualitativa, fundada na extração textual dos dados do modelo do perfil. Desta forma, é no modelo de perfil de pesquisadores que se buscam os

dados e, é no modelo de similaridade do perfil de pesquisadores que se observa como encontrar as similaridades.

3.3.1. Similaridade Baseado na Produção Quantitativa

Na similaridade ponderada do perfil do pesquisador com base na sua produção quantitativa, o seu resultante é um número entre 0 e 1, sendo que quanto mais perto de 1 mais similar e quanto mais perto de 0 menos similar. A similaridade local pelo elemento do perfil utiliza o valor do elemento do modelo de perfil de dois pesquisadores, multiplicado pelo peso do elemento, onde o valor do peso do elemento é dividido pelo total de peso do modelo, indicando a importância do elemento para determinar a similaridade, conforme é demonstrado na Equação (1).

$$sim\_quanti\_e_{(pa,pb)} = \frac{min(e_{pa}, e_{pb})}{max(e_{pa}, e_{pb})}.p \quad (1)$$

Sendo,

- pa = pesquisador a;
- pb = pesquisador b;
- e = valor do elemento do perfil do pesquisador;
- $min(e_{pa}, e_{pb})$ = representa o menor valor do elemento entre pa e pb;
- $max(e_{pa}, e_{pb})$ = representa o maior valor do elemento entre pa e pb;
- p = peso do elemento.

Visando uma melhor compreensão da métrica proposta, a Tabela 6 indica um exemplo do cálculo de um elemento do modelo de perfil de pesquisadores tendo um peso igual a 12 para o elemento, sendo este dividido pelo total dos pesos dos elementos. Assim, o pesquisador A tem um quantitativo de 13 artigos publicados e o pesquisador B, 15 artigos publicados. Com isso, aplica-se a Equação (1) e obtém-se o valor da similaridade local deste elemento. Pode-se observar que, para esse elemento, terá no máximo um valor de similaridade de 0,12 a ser somado aos outros elementos do perfil.

Tabela 6. Exemplo de cálculo de similaridade local do elemento quantitativo.

Elementos	**Sigla**	**Peso**	**Máx. Valor Intervalo**	**Pesquisador A**	**Pesquisador B**	***sim_quanti_e*$_{(pa,pb)}$**
Artigos em Periódicos	AP	12	96	**13**	**15**	0,10
Valor da similaridade local do elemento Artigos em Periódicos: (13/15) * (12/100) = **0,10**						

Na similaridade local pela categoria do perfil, apresenta-se o somatório das medidas de similaridades locais, ponderando cada elemento da categoria do modelo do perfil, conforme a Equação (2).

$$sim\_quanti\_c_{(pa,pb)} = \sum_{i=1}^{c} \frac{min(e_{i,pa}\ , e_{i,pb})}{max(e_{i,pa}\ , e_{i,pb})} . p_i \qquad (2)$$

Sendo,

- pa = pesquisador a;
- pb = pesquisador b;
- c = representa o número total de elementos da categoria do modelo de perfil;
- i = intervalo de 1 até o número total de elementos da categoria c;
- e_i = valor do elemento do perfil do pesquisador no intervalo i;
- $min(e_{i,pa}\ , e_{i,pb})$ = representa o menor valor do elemento entre $e_{i,pa}$ e $e_{i,pb}$;
- $\max(e_{i,pa}\ , e_{i,pb})$ = representa o maior valor do elemento entre $e_{i,pa}$ e $e_{i,pb}$;
- p_i = peso do elemento no intervalo i.

Para uma melhor compreensão da métrica, um exemplo é apresentado na Tabela 7, onde é demonstrado o cálculo local da categoria do modelo de perfil de pesquisadores. Informando-se os valores dos elementos do pesquisador A, juntamente com os valores quantitativos do pesquisador B, é feito o cálculo de cada elemento e, somando-se pelos elementos da categoria em questão, conforme a Equação (2).

Tabela 7. Exemplo de cálculo de similaridade local da categoria quantitativo.

Categoria	Peso	Elementos	Peso	PA	PB	$sim\_quanti\_e_{(pa,pg)}$	$sim\_quanti\_c_{(pa,pb)}$
Produção	40	Artigos em Periódicos	12	**15**	**20**	0,09	0,31
		Capítulos de Livros	3	**10**	**15**	0,02	
		Livros	5	**5**	**4**	0,04	
		Trabalhos Completo em Conferências	8	**20**	**15**	0,06	
		Rede de Coautoria	4	**19**	**19**	0,04	
		Projeto de Pesquisas	5	**10**	**10**	0,05	
		Softwares	3	**6**	**2**	0,01	
Valor da similaridade local da categoria quantitativo: (15/20) * (12/100) + (10/15) * (3/100) + (4/5) * (5/100) + (15/20) * (8/100) + (19/19) * (4/100) + (10/10) * (5/100) + (2/6) * (3/100) = **0,31**							

Para a similaridade global do perfil, calcula-se o somatório das medidas de similaridades locais, ponderando cada elemento de todos os elementos do modelo do perfil, conforme a Equação (3). O exemplo de cálculo é igual ao da Tabela 7, porém, se estende para todos os elementos do perfil.

$$sim\_quanti\_g_{(pa,pb)} = \sum_{i=1}^{g} \frac{min(e_{i,pa}, e_{i,pb})}{max(e_{i,pa}, e_{i,pb})} . p_i \quad (3)$$

Sendo,

- pa = pesquisador a;
- pb = pesquisador b;
- g = representa o número total de elementos do modelo de perfil;
- i = intervalo de 1 até o número total de elementos do perfil g;
- e_i = valor do elemento do perfil do pesquisador no intervalo i;
- $min(e_{i,pa}, e_{i,pb})$ = representa o menor valor do elemento entre $e_{i,pa}$ e $e_{i,pb}$;
- $max(e_{i,pa}, e_{i,pb})$ = representa o maior valor do elemento entre $e_{i,pa}$ e $e_{i,pb}$;
- p_i = peso do elemento no intervalo i.

O índice de similaridade quantitativo demonstra o quão similares são dois perfis de pesquisadores conforme o modelo, dentro das características referentes a seus estudos e trabalhos. Isto é realizado medindo a similaridade do quanto foi produzido durante a vida acadêmica do pesquisador.

3.3.2. Similaridade Baseada na Produção Qualitativa

Na similaridade baseada na produção qualitativa, a extração textual dos dados do perfil é modelada como um vetor de termos/características, onde os dados são retirados dos elementos do modelo de perfil. O objetivo, na fase inicial, é transformar os dados dos atributos da produção do pesquisador em números de forma significativa e, assim, compreender melhor o perfil. Desta forma, os mesmos podem ser calculados e comparados a outros perfis.

Nesse contexto, é possível observar a frequência de ocorrência do termo no elemento do perfil, isto é, o número de vezes que o termo t aparece no elemento do perfil p, como também o total da frequência do termo no perfil, que é a soma das frequências de ocorrência do termo t em todos os elementos e do perfil p. Destaca-se que a frequência do termo é o número de vezes que o termo t ocorre no perfil do pesquisador, e para cada elemento faz-se a ponderação do mesmo. Nesse caso, o cálculo da frequência ponderada de ocorrência do termo t nos elementos do perfil do pesquisador pa é feito pelo somatório da frequência do termo em cada elemento do perfil de forma ponderada, conforme a Equação (4).

$$fp_{(t,pa)} = \sum_{i=1}^{e} ft_i . p_i \qquad (4)$$

Sendo,

- $fp_{(t,pa)}$ = frequência ponderada do termo t no perfil do pesquisador pa;
- e = representa o número total de elementos modelo de perfil;
- i = representa o intervalo de 1 até o número total de elementos e;
- ft_i = representa a frequência do termo no elemento no intervalo de i;
- p_i = peso do elemento no intervalo de i.

Objetivando uma melhor compreensão da métrica proposta, a Tabela 8 demonstra um exemplo do cálculo da frequência ponderada do termo do modelo de perfil de pesquisadores. Informando-se os valores da quantidade de vezes que o termo ocorreu nos elementos do perfil do pesquisador A, e é feito o cálculo de cada elemento, somando-se por todos os elementos em questão, conforme a Equação (4).

Tabela 8. Exemplo de cálculo da frequência ponderada do termo.

Categoria	**Peso**	**Elementos**	**Peso**	**Termo = Banco de Dados**	$fp_{(t,pa)}$
Produção	40	Artigos em Periódicos	12	**10**	3,11
		Capítulos de Livros	3	**2**	
		Livros	5	**0**	
		Trabalhos Completo em Conferências	8	**20**	
		Rede de Coautoria	4	**0**	
		Projeto de Pesquisas	5	**5**	
		Softwares	3	**0**	
Valor da frequência ponderada do termo Banco de Dados: (10 * (12/100)) + (2 * (3/100)) + (0 * (5/100)) + (20 * (8/100)) + (0 * (4/100)) + (5 * (5/100)) + (0 * (3/100)) = **3,11**					

A frequência ponderada do termo demonstra a quantidade de vezes que o termo aparece no perfil do pesquisador conforme o modelo, e o pondera. Isto possibilita que seja colocada, uma importância para tal termo, o qual ainda deve passar por uma verificação da sua relevância com o passar do tempo da vida acadêmica do pesquisador.

3.3.3. Realimentação da Relevância do Termo para o Perfil dos Pesquisadores

A produção dos pesquisadores é contínua e o currículo, baseado em sua vida acadêmica, pode expressar mudanças no comportamento de seus elementos do perfil com o passar dos anos. Dessa forma, alguns termos antes importantes para a construção de suas atividades, podem, posteriormente, serem substituídos por novos. Essas mudanças impactam diretamente em suas similaridades e, consequentemente, em seus interesses. Desta forma, um termo que, mesmo bem ponderado, pode não refletir a necessidade de recomendação atual de novos conteúdos. Como diferentes termos têm diferentes graus de importância para fins descritivos do perfil, de que forma é possível verificar tal importância? Isso pode ser feito

com a atribuição de peso numérico a cada termo que descreve o perfil. Desta forma, a fim de caracterizar a importância dos termos para cada perfil dos pesquisadores, é calculado um grau de relevância $gr_{(t,pa)}$ sendo associado a cada termo t de um perfil de pesquisador pa. A realimentação dessa relevância deve ser constante, pois o ambiente em que está inserido é mutável, com muitas adaptações, o que indica a necessidade de um mecanismo que realimente de tempos em tempos a relevância do termo para o perfil.

O grau de relevância quantifica a importância do termo na descrição do perfil. Para identificar o grau de relevância de um termo busca-se o somatório dos intervalos de anos do currículo do pesquisador e o somatório de todos os elementos do perfil, com a frequência em que o termo ocorre dentro do elemento em determinado ano. Devido a estas características temporais, multiplica-se o resultado anterior pelo peso do elemento e divide-se pelo intervalo do ano. Desta forma, um termo atual, não incorporado ao perfil em suas próximas atividades, tende a diminuir de relevância para o perfil e, deve perder pontos dentro do modelo de similaridade.

A similaridade dos perfis ocorre pela comparação do conjunto de termos relacionados à sua produção, sendo que os termos terão um grau de relevância medido através de frequência do termo ponderado e em relação ao ano em que o termo foi utilizado na produção. O cálculo do grau de relevância dos termos nos elementos do perfil é calculado, pelo somatório dos intervalos de anos do currículo do pesquisado, do somatório de todos os elementos do perfil com a frequência que o termo ocorre dentro do elemento em determinado ano, multiplicando-se o resultado anterior pelo peso do elemento. Por fim, divide-se pelo intervalo do ano, conforme a Equação (5).

$$gr_{(t,pa)} = \sum_{j=1}^{a} \frac{\sum_{i=1}^{e} ft_{i,j}.p_i}{j} \quad (5)$$

Sendo,

- $gr_{(t,pa)}$ = grau de relevância de um termo t extraído dos elementos do perfil do pesquisador pa;
- a = representa o número total anos da produção do pesquisador;
- j = representa o intervalo de 1 até o número total de a. Sendo que 1 representa o ano atual, 2 representa o ano anterior a 1, seguindo a sequência até o total de a;
- e = representa o número total de elementos do modelo de perfil;

- i = representa o intervalo de 1 até o número total de elementos e;
- $ft_{i,j}$ = representa a frequência do termo no intervalo do elemento i, dentro do intervalo de ano j;
- p_i = peso do elemento no intervalo de i.

Para uma melhor compreensão da métrica proposta, um exemplo é apresentado na Tabela 9, onde é demonstrado o cálculo de um termo retirado do perfil do pesquisador dentro de um dos elementos, com intervalo temporal entre o ano atual até o ano final encontrado no currículo do pesquisador. Dentro do ano, é verificado a quantidade de vezes que esse termo ocorreu. Após, é aplicada a Equação (5), lembrando que esse termo é verificado em cada elemento do modelo de perfil de pesquisadores e, é feita uma soma dos resultados. Esse processo é realizado para cada termo encontrado nos dados do pesquisador. Conforme passam-se os anos, a realimentação dos termos é refeita ao se calcular novamente o grau de relevância e, se o termo não estiver nos dados do pesquisador, o grau de relevância tende a diminuir, conforme se observa no cálculo de exemplo da

Tabela 10.

Tabela 9. Exemplo de cálculo do grau de relevância do termo no ano atual.

Elementos	**Sigla**	**Peso**	**Elemento Textual = Termo**	**Pesquisador**	**2017**	**2016**	**2015**	$gr_{(t,pa)}$
Artigos em Periódicos	AP	12	Banco de Dados	**A**	**15**	**10**	**5**	2,6
Valor do grau de relevância do termo no elemento Artigos em Periódicos: (15 * (12/100)) / 1 + (10 * (12/100)) / 2 + (5 * (12/100)) / 3 = 1,8 + 0,6 + 0,2 = **2,6**								

Tabela 10. Exemplo de cálculo do grau de relevância do termo no ano seguinte.

Elementos	**Sigla**	**Peso**	**Elemento Textual = Termo**	**Pesquisador**	**2018**	**2017**	**2016**	**2015**	$gr_{(t,pa)}$
Artigos em Periódicos	AP	12	Banco de Dados	**A**	**0**	**15**	**10**	**5**	1,45
Valor do grau de relevância do termo no elemento Artigos em Periódicos: (0 * (12/100)) / 1 + (15 * (12/100)) / 2 + (10 * (12/100)) / 3 + (5 * (12/100)) / 4 = 0 + 0,9 + 0,4 + 0,15 = **1,45**									

Desta forma o grau de relevância de um termo descreve numericamente as preferências atualizadas do perfil do pesquisador.

3.3.4. Índice de Similaridade Qualitativa

O grau de relevância resulta em um inteiro positivo, o qual tem a função de verificar dentro do próprio perfil o que é mais relevante para o pesquisador. Desta forma, busca-se dentro de cada perfil os termos com o maior grau de relevância. Assim, é montado um vetor com esses termos, comparando com os de outros pesquisadores, por meio da Equação (6), a qual descreve o índice de similaridade Jaccard [46] que mede o grau de similaridade entre dois conjuntos de elementos.

$$sim\_vt_{(pa,pb)} = \frac{tc_{(pa,pb)}}{tc_{(pa,pb)} + te_{(pa)} + te_{(pb)}} \quad (6)$$

Sendo,

- $sim\_vt_{(pa,pb)}$ = similaridade entre o vetor de termos do pesquisador pa e o pesquisador pb;
- $tc_{(pa,pb)}$ = número de termos em comum entre os dois vetores;
- $te_{(pa)}$ = número de termos exclusivos do vetor de termos do pesquisador pa;
- $te_{(pb)}$ = número de termos exclusivos do vetor de termos do pesquisador pb.

Objetivando uma melhor compreensão da métrica proposta, na Tabela 11 é demonstrado o cálculo da similaridade entre os dois vetores de termos dos pesquisadores utilizando a Equação (6).

Tabela 11. Exemplo de cálculo da similaridade dos vetores de termos.

Elementos Textuais	**Pesquisador A**	**Pesquisador B**	$sim\_vt_{(pa,pb)}$
Termos com maior grau de relevância	**banco de dados, java, algoritmos, análise de imagens, grafos**	**php, reconhecimento de padrões, banco de dados, algoritmos, grafos**	0,42
Valor da similaridade qualitativa verificando o os termos mais relevantes do perfil: 3 / (3 + 2 + 2) = **0,42**			

Após encontrar os termos em comum entre os dois vetores, calcula-se a similaridade entre o grau de relevância dos termos em comum. Nesse cálculo, a similaridade do grau de relevância de um termo *t*, entre o pesquisador A e B, será a divisão do menor grau de relevância entre A e B pelo maior. Assim, essas similaridades entre os termos são somadas e divididas pelo número de termos em comum, conforme a Equação (7).

$$sim\_tc_{(pa,pb)} = \frac{\sum_{i=1}^{tt} \frac{min(gr_{(t,pa)i}, gr_{(t,pb)i})}{max(gr_{(t,pa)i}, gr_{(t,pb)i})}}{tt} \quad (7)$$

Sendo,

- pa = pesquisador a;
- pb = pesquisador b;
- tt = representa o número total de termos em comum entre os pesquisadores;
- i = intervalo de 1 até o número total de termos em comum tt;
- $gr_{(t,pa)i}$ = valor do grau de relevância do termo *t* do perfil do pesquisador no intervalo i;
- $min(gr_{(t,pa)i}, gr_{(t,pb)i})$ = representa o menor valor do grau de relevância entre $gr_{(t,pa)i}$ e $gr_{(t,pb)i}$;
- $max(gr_{(t,pa)i}, gr_{(t,pb)i})$ = representa o maior valor do grau de relevância entre $gr_{(t,pa)i}$ e $gr_{(t,pb)i}$.

Para compreender melhor a métrica proposta, a Tabela 12 demonstra um cálculo da similaridade entre os termos em comum, conforme a Equação (7).

Tabela 12. Exemplo de cálculo da similaridade entre os graus de relevância dos termos em comum aos perfis.

Elementos Textuais	$gr_{(t,pa)}$	$gr_{(t,pb)}$	$sim\_gr(gr_{(t,pa)}, gr_{(t,pb)})$	$sim\_tc_{(pa,pb)}$
Banco de Dados	**5**	**0,92**	0,18	
Algoritmos	**0,50**	**0,50**	1	0,46
Grafos	**0,68**	**3**	0,22	
Valor da similaridade entre os termos mais relevantes em comum dos perfis: ((0,92/5) + (0,50/0,50) + (0,68/3)) / 3 = **0,46**				

Por fim, o índice de similaridade qualitativa entre o perfil dos pesquisadores é dado pela soma das similaridades do vetor de termos, junto com a similaridade entre os graus de relevância dos termos em comuns dos pesquisadores, dividido por 2, conforme a Equação (8).

$$sim\_quali_{(pa,pb)} = \frac{sim\_vt_{(pa,pb)} + sim\_tc_{(pa,pb)}}{2} \quad (8)$$

Sendo,

- pa = pesquisador a;
- pb = pesquisador b;
- $sim\_vt_{(pa,pb)}$ = representa o número da similaridade do vetor de termos entre o pesquisador pa e pb;
- $sim\_tc_{(pa,pb)}$ = representa o número da similaridade dos termos em comum entre o pesquisador pa e pb.

Um exemplo de cálculo da similaridade qualitativa pode ser visualizado na Tabela 13, conforme Equação (8).

Tabela 13. Exemplo de cálculo da similaridade qualitativa.

$sim\_vt_{(pa,pb)}$	$sim\_tc_{(pa,\ pb)}$	$sim\_quali_{(pa,pb)}$
0,42	**0,46**	0,44
Valor da similaridade qualitativa: (0,42 + 0,46) / 2 = **0,44**		

O índice de similaridade qualitativa demonstra o quão similares são dois perfis de pesquisadores conforme o modelo, dentro das características referentes a seus estudos e trabalhos. Isto é realizado medindo a similaridade do que foi produzido durante a vida acadêmica do pesquisador.

3.4. MÉTRICAS DE RECOMENDAÇÃO

Com a estratégia de recomendação baseada exclusivamente em conteúdo, com fundamento na vida acadêmica do pesquisador, e em seu comportamento, através de suas atividades acadêmicas, são medidas as relevâncias nos itens de seus elementos do modelo de perfil. Desta forma, o algoritmo “recomendador” utiliza os índices similaridades quanti e/ou

quali, global e/ou local, juntamente com o grau de relevância dos termos, para identificar itens de perfis similares e os próprios perfis que sejam compatíveis com o do perfil indicado para a recomendação acontecer.

3.4.1. Regras e Funções de Recomendação

Para as recomendações, definiu-se algumas regras, respondendo as seguintes questões: O que recomendar? Como recomendar? Em busca das respostas, definiu-se as funções de recomendação, nas quais são comparados pares de pesquisadores pa e pb, e se há satisfações às regras, a recomendação ocorrerá. Para a comparação das similaridades, adotou-se as seguintes interpretações aos valores obtidos dos cálculos de similaridade, conforme a Tabela 14 adaptado da medida de similaridade de Coeficiente de Pearson.

Tabela 14. Correlação dos valores das medidas de similaridade.

Valor	**Interpretação**
$p \geq 0{,}7 \wedge \leq 1$	Forte
$p \geq 0{,}4 \wedge < 0{,}7$	Moderada
$p \geq 0{,}2 \wedge < 0{,}4$	Fraca
$p \geq 0 \wedge < 0{,}2$	Ausência

A seguir é apresentado as recomendações e suas regras e funções.

Recomendação 1: Recomendação de outro pesquisador.

O que recomendar: O perfil de outros pesquisadores para networking.

Como recomendar: Baseado na similaridade global do modelo.

Exemplo: O pesquisador receberá indicação dos perfis mais semelhantes ao dele levando em consideração todos os elementos do perfil.

A Equação (9) apresenta a ação de recomendação: Pesquisador para networking. Se a similaridade quantitativa global e a similaridade qualitativa entre o pesquisador pa e pb pertencerem ao intervalo forte ou moderado.

$$r_{(pa,pb)} = \begin{cases} recomenda\ pesquisador\ para\ networking, \\ se\ (sim\_quanti\_g_{(pa,pb)} \in \{Forte, Moderado\}) \wedge \\ (sim\_quali_{(pa,pb)} \in \{Forte, Moderado\}); \end{cases} \quad (9)$$

Sendo,

- r = recomendação;
- pa = pesquisador a;
- pb = pesquisador b;
- sim_quanti_g = similaridade global do perfil, ponderada e baseada na produção quantitativa;
- sim_quali = similaridade global do perfil, ponderada e baseada na extração textual, produção qualitativa.

Recomendação 2: Recomendação de pesquisador para orientação em conjunto.

O que recomendar: Pesquisadores para parceria em orientações de mestrado, doutorado e pós-doutorado.

Como recomendar: Processar os elementos da categoria orientação e o elemento de formação acadêmica do pesquisador. Medir a similaridade dos elementos da categoria de orientação e o índice de similaridade qualitativa.

Exemplo: Pesquisador que realiza orientações em mestrado, doutorado ou pós-doutorado receberá recomendações de outros perfis com essas similaridades, ou seja, que também costumam realizar orientações.

A Equação (10) apresenta as ações de recomendação: Parceiro para orientação de mestrado, de doutorado e/ou pós-doutorado. Se os elementos om, od, op dos pesquisadores pa e pb, forem maior que zero, e se a formação acadêmica de pa e de pb pertencer a doutorado ou pós-doutorado. E ainda, se a similaridade quantitativa dos elementos e a similaridade qualitativa entre o pesquisador pa e pb, pertencerem ao intervalo forte ou moderado.

$$r_{(pa,pb)} = \begin{cases} \begin{array}{c} recomenda\ parceiro\ para\ orientação\ de\ mestrado, \\ se\ (om_{pa} > 0) \wedge (om_{pb} > 0) \wedge \\ (fa_{pa} \wedge fa_{pb} \in \{Doutorado, Pós - Doutorado\}) \wedge \\ (sim\_quanti\_e_{(om_{pa}, om_{pb})} \in \{Forte, Moderado\}) \wedge \\ (sim\_quali_{(pa,pb)} \in \{Forte, Moderado\}); \end{array} \\ \\ \begin{array}{c} recomenda\ parceiro\ para\ orientação\ de\ doutorado, \\ se\ (od_{pa} > 0) \wedge (od_{pb} > 0) \wedge \\ (fa_{pa} \wedge fa_{pb} \in \{Doutorado, Pós - Doutorado\}) \wedge \\ (sim\_quanti\_e_{(od_{pa}, od_{pb})} \in \{Forte, Moderado\}) \wedge \\ (sim\_quali_{(pa,pb)} \in \{Forte, Moderado\}); \end{array} \\ \\ \begin{array}{c} recomenda\ parceiro\ para\ orientação\ de\ pós - doutorado, \\ se\ (op_{pa} > 0) \wedge (op_{pb} > 0) \wedge \\ (fa_{pa} \wedge fa_{pb} \in \{Pós - Doutorado\}) \wedge \\ (sim\_quanti\_e_{(op_{pa}, op_{pb})} \in \{Forte, Moderado\}) \wedge \\ (sim\_quali_{(pa,pb)} \in \{Forte, Moderado\}); \end{array} \end{cases} \quad (10)$$

Sendo,

- r = recomendação;
- pa = pesquisador a;
- pb = pesquisador b;
- om = elemento do perfil referente a orientações de mestrado;
- od = elemento do perfil referente a orientações de doutorado;
- op = elemento do perfil referente a orientações de pós-doutorado;
- fa = elemento do perfil referente a formação acadêmica;
- sim_quanti_e = similaridade local do elemento da categoria de orientação, ponderada e baseada na produção quantitativa;
- sim_quali = similaridade global do perfil, ponderada e baseada na extração textual, produção qualitativa.

Recomendação 3: Recomendação de membros para bancas de avaliação.

O que recomendar: Pesquisadores para participação em bancas de mestrado e doutorado.

Como recomendar: Verificar a similaridade local da categoria do pesquisador, sua formação acadêmica, sua área de conhecimento, juntamente com suas participações em bancas.

Exemplo: Pesquisador que costuma participar em bancas de mestrado ou doutorado receberá recomendações de outros pesquisadores da mesma área do conhecimento que também participem de bancas.

A Equação (11) apresenta as ações de recomendação: Participantes para banca de mestrado e/ou doutorado. Se a área do conhecimento do pesquisador pa e pb forem as mesmas, se os elementos pbm/pbd de pa e pb forem maior que zero, e se a formação acadêmica de pa e de pb pertencer à doutorado ou pós-doutorado. E ainda, se a similaridade quantitativa da categoria e a similaridade qualitativa entre o pesquisador pa e pb, pertencerem ao intervalo forte ou moderado.

$$r_{(pa,pb)} = \begin{cases} recomenda\ participantes\ para\ bancas\ de\ mestrado, \\ se\ (ac_{pa} = ac_{pb}) \wedge (pbm_{pa} > 0) \wedge (pbm_{pb} > 0) \wedge \\ (fa_{pa} \wedge fa_{pb} \in \{Doutorado, Pós - Doutorado\}) \wedge \\ (sim\_quanti\_c_{(ban_{pa}, ban_{pb})} \in \{Forte, Moderado\}) \wedge \\ (sim\_quali_{(pa,pb)} \in \{Forte, Moderado\}); \\ \\ recomenda\ participantes\ para\ bancas\ de\ doutorado, \\ se\ (ac_{pa} = ac_{pb}) \wedge (pbd_{pa} > 0) \wedge (pbd_{pb} > 0) \wedge \\ (fa_{pa} \wedge fa_{pb} \in \{Doutorado, Pós - Doutorado\}) \wedge \\ (sim\_quanti\_c_{(ban_{pa}, ban_{pb})} \in \{Forte, Moderado\}) \wedge \\ (sim\_quali_{(pa,pb)} \in \{Forte, Moderado\}); \end{cases} \quad (11)$$

Sendo,

- r = recomendação;
- pa = pesquisador a;
- pb = pesquisador b;
- pbm = elemento do perfil referente a participação em bancas de mestrado;
- pbd = elemento do perfil referente a participação em bancas de doutorado;
- fa = elemento do perfil referente a formação acadêmica;
- sim_quanti_c = similaridade local da categoria banca, ponderada e baseada na produção quantitativa;

- sim_quali = similaridade global do perfil, ponderada e baseada na extração textual, produção qualitativa.

Recomendação 4: Recomendação de orientador para futura formação do pesquisador.

O que recomendar: Recomendação de orientador para futura formação do pesquisador.

Como recomendar: Verificar na categoria do pesquisador sua formação acadêmica e áreas de conhecimento. Encontrar pesquisador semelhante com formação superior e que realize orientações na formação pretendida.

Exemplo: Pesquisador com mestrado na área de educação, receberá recomendações de outros pesquisadores com formação de doutorado e que realizam orientações de doutorado nesta área.

A Equação (12) apresenta as ações de recomendação: Possível orientador Doutorado/Pós-Doutorado para futura formação. Se a área do conhecimento do pesquisador pa e pb forem as mesmas, se os elementos od/op de pb for maior que zero. Se a formação acadêmica de pa for igual a mestrado/doutorado, então a formação de pb deve pertencer à doutorado ou pós-doutorado. E ainda, se a similaridade qualitativa entre o pesquisador pa e pb, pertencerem ao intervalo forte ou moderado.

$$r_{(pa,pb)} = \begin{cases} recomenda\ orientador\ para\ futura\ formação\ de\ doutorado, \\ se\ (ac_{pa} = ac_{pb}) \wedge (od_{pb} > 0) \wedge \\ (fa_{pa} = Mestrado \rightarrow fa_{pb} \in \{Doutorado, Pós - Doutorado\}) \wedge \\ (sim\_quali_{(pa,pb)} \in \{Forte, Moderado\}); \\ \\ recomenda\ orientador\ para\ futura\ formação\ de\ Pós - doutorado, \\ se\ (ac_{pa} = ac_{pb}) \wedge (op_{pb} > 0) \wedge \\ (fa_{pa} = Doutorado \rightarrow fa_{pb} \in \{Pós - Doutorado\}) \wedge \\ (sim\_quali_{(pa,pb)} \in \{Forte, Moderado\}); \end{cases} \quad (12)$$

Sendo,

- r = recomendação;
- pa = pesquisador a;
- pb = pesquisador a;
- ac = elemento do perfil referente a área do conhecimento;
- od = elemento do perfil referente a orientações de doutorado;

- op = elemento do perfil referente a orientações de pós-doutorado;
- fa = elemento do perfil referente a formação acadêmica;
- sim_quali = similaridade global do perfil, ponderada e baseada na extração textual, produção qualitativa.

Recomendação 5: Recomendação de produções de outros pesquisadores.

O que recomendar: Publicações como artigos, livros, capítulo de livros e trabalhos completos em conferências.

Como recomendar: Baseado na similaridade qualitativa dos perfis, onde a área do conhecimento seja compatível, sendo que o ano da publicação seja recente e os termos mais relevantes do primeiro perfil devem conter os termos do outro perfil.

Exemplo: Recomendar publicações de pesquisador com perfil semelhante, onde as produções são dos últimos 5 anos. Da mesma forma que, os cinco principais termos do pesquisador B, os quais tenham o maior grau de relevância, estejam contido no perfil do pesquisador A.

A Equação (13) apresenta a ação de recomendação: Artigo em periódico, capítulo de livro, livro, trabalho completo em evento. Se a área de conhecimento do pesquisador A for compatível com a área de conhecimento do pesquisador B; se o ano da produção for maior ou igual ao ano corrente menos 5; se os termos que descrevem o perfil de pa contém os termos com maior grau de relevância de pb; se a similaridade qualitativa entre o pesquisador pa e pb pertencerem ao intervalo forte ou moderado.

$$r_{(pa,pb)} = \begin{cases} recomenda\ artigo\ em\ periódico, \\ capítulo\ de\ livro, \\ livro, \\ trabalho\ completo\ em\ evento, \\ se\ (ac_{pa} = ac_{pb}) \wedge \\ (ano_{produção} \geq (ano_{corrente} - 5)) \wedge \\ (termos_{pa} \supset termos_{pb}) \wedge \\ (sim\_quali_{(pa,pb)} \in \{Forte, Moderado\}); \end{cases} \quad (13)$$

Sendo,

- r = recomendação;

- pa = pesquisador a;
- pb = pesquisador b;
- ac = elemento do perfil referente a área do conhecimento;
- $ano_{corrente}$ = ano corrente do cálculo das recomendações, diminui-se 5 deste para verificar apenas a produção mais recente do pesquisador;
- $ano_{produção}$ = ano em que a produção foi publicada;
- $termos_{pa}$ = os termos do pesquisador;
- $termos_{pb}$ = os 5 termos do pesquisador com maior $GR_{(t,pb)}$;
- sim_quali = similaridade global do perfil, ponderada e baseada na extração textual, produção qualitativa.

Recomendação 6: Recomendação de periódicos para o pesquisador.

O que recomendar: ISSN, nome, área e estrato do periódico para o pesquisador.

Como recomendar: Verificar se os pesquisadores realizam publicações em periódicos e recomendar os periódicos que os perfis semelhantes publicam.

Exemplo: Pesquisador A tem um perfil semelhante ao perfil do pesquisador B, então periódicos em que o pesquisador *B* publica serão recomendados para o pesquisador *A* publicar também.

A Equação (14) apresenta a ação de recomendação: Periódico. Se o elemento artigo em periódico do pesquisador A e B for maior que zero; se a área de conhecimento do pesquisador A pertencer a mesma área de avaliação do periódico em que o pesquisador B publicou; se a similaridade quantitativa global e a similaridade qualitativa entre o pesquisador pa e pb pertencerem ao intervalo forte ou moderado.

$$r_{(pa,pb)} = \begin{cases} recomenda\ periódico, \\ se\ (ap_{pa} > 0) \wedge (ap_{pb} > 0\ ) \wedge \\ \left(ac_{pa} \in \left\{area_{periodico_{pb}}\right\}\right) \wedge \\ (sim\_quanti\_g_{(pa,pb)} \in \{Forte, Moderado\}) \wedge \\ (sim\_quali_{(pa,pb)} \in \{Forte, Moderado\}); \end{cases} \quad (14)$$

Sendo,

- r = recomendação;

- pa = pesquisador a;
- pb = pesquisador b;
- ap = elemento do perfil referente a artigos em periódicos;
- $area_{periodico}$ = área de avaliação do periódico pela CAPES;
- ac = elemento do perfil referente a área do conhecimento;
- sim_quanti_g = similaridade global do perfil, ponderada e baseada na produção quantitativa;
- sim_quali = similaridade global do perfil, ponderada e baseada na extração textual, produção qualitativa.

3.5. PROCESSO DE DELIMITAÇÃO, COLETA E PREPARAÇÃO DOS DADOS

Para uma melhor visualização de todo o processo de coleta do dados, da delimitação da quantidade de dados a serem utilizados e a preparação dos mesmo, dividiu-se o processo em etapas, no qual evidencia-se dez passos. Então, nesses passos foram coletados os dados necessários para compor o perfil do pesquisador, levado em consideração o modelo de perfil adaptado Rep-Model. Em seguida, aplicou-se os cálculos referente a abordagem proposta gerando assim as similaridades entre os perfis e suas recomendações, para então se desenvolveu uma ferramenta/aplicação que tem por objetivo o acesso aos dados processados. Posteriormente, são feitos experimento junto aos dados para validar a abordagem proposta. A Figura 9 demonstra de forma mais intuitiva todo o processo descrito até aqui.

Figura 9: Visualização do processo de delimitação, coleta, preparação e cálculo dos dados.

O CNPq disponibiliza, periodicamente, todos os números identificadores dos Currículos Lattes cadastrados em sua plataforma no formato aberto[18]. Os arquivos foram baixados em janeiro de 2017, com os nomes numero_identificador_lattes_20170108.csv, tab_area_conhecimento_20170108.csv e tab_nivel_formacao_20170108.csv.

O arquivo numero_identificador_lattes_20170108.csv contém aproximadamente 5 milhões de linhas, as quais representam todos os currículos cadastrados na base da Plataforma Lattes. Cada linha contém o número identificador do currículo, a nacionalidade, a data de atualização do currículo, o código da área do conhecimento e o código do nível de formação. Destes, somente o identificador é garantido estar presente no arquivo. O arquivo tab_area_conhecimento_20170108.csv contém os dados das áreas do conhecimento e o arquivo tab_nivel_formacao_20170108.csv contém os dados dos níveis de formação da Plataforma Lattes. Os dados desses arquivos foram importados para uma base de dados relacional para melhor verificação, exploração e seleção dos mesmos.

Com base nesses arquivos, delimitou-se os dados a serem utilizados e, de forma aleatória, buscou-se selecionar 8 áreas do conhecimento de um total de 86 áreas do arquivo tab_area_conhecimento_20170108.cvs, tendo o cuidado de ser uma área de cada grande área. A Tabela 15 mostra as áreas selecionadas para o uso. Dentro das áreas selecionadas, buscou-se os identificadores dos currículos da plataforma contidos no arquivo numero_identificador_lattes_20170108.cvs. Selecionou-se todos com os seguintes níveis de formação: Mestrado, Mestrado Profissional, Doutorado, Pós-Doutorado e Livre Docência. Como resposta obteve-se 157.916 registros.

Tabela 15. Áreas selecionadas para coleta de dados.

Grande Área	**Área**
Ciências Exatas e da Terra	Ciência da Computação
Ciências Biológicas	Genética
Engenharias	Engenharia Elétrica
Ciências da Saúde	Odontologia
Ciências Agrárias	Agronomia
Ciências Sociais Aplicadas	Economia
Ciências Humanas	Educação
Linguística	Letras

[18] Disponível em http://memoria.cnpq.br/web/portal-lattes/extracoes-de-dados.

Com os registros selecionados, buscou-se estudo e formas de extrair os dados dos currículos da plataforma, em formato XML, de forma automatizada. Buscou-se no site da Plataforma Lattes por um arquivo chamado robots.txt e o mesmo não foi encontrado. Tal arquivo, contém dados de permissão ou negação de indexação dos dados do site, assim, construiu-se um script para baixar os dados do currículo de forma automática. A extração foi realizada em duas etapas: (i) na primeira, buscou correlacionar os números identificadores da plataforma, pré-selecionados na etapa de delimitação dos dados a serem utilizados, com o número identificador "menor" utilizado para a consulta do currículo na web; (ii) feita a correlação, a próxima etapa foi realizar o download dos arquivos XMLs dos currículos, os quais foram baixados de forma compactada. Ressalta-se que o script correlacionou e baixou, um total de 157.601 arquivos, sendo esse o total de currículos utilizados.

Na sequência das atividades, os arquivos foram descompactados, criando-se uma base de dados junto ao software BaseX, com o intuito de estudar os dados dos XMLs, juntamente com os arquivos DTD e XSD. Assim, foram identificados os endereçamentos dos dados necessários para a criação do perfil, ou seja, preencher o modelo de perfil e seus elementos. Usando consultas xQuery, foram extraídos conjuntos de dados para realizar a carga dos mesmos, numa base de dados relacional. Todo esse processo pode ser melhor visualizado no Apêndice A.

Depois de realizar o estudo dos dados dos arquivos XMLs, identificou-se os dados para compor os perfis dos pesquisadores, de acordo com o modelo adaptado do Rep-Model. Dessa forma, foram criadas as tabelas para a importação dos mesmos, utilizando-se o banco de dados PostgreSQL para esta tarefa.

Depois de criadas as tabelas de acordo com os dados vindos do XML, foi estudada a melhor forma para realizar a carga para dentro da base de dados relacional. Assim, escolheu-se o software *Pentaho Data Integration*[19] para realização da tarefa. Trata-se esse de um software de ETL (*Extract-Transform-Load*), com objetivo de extração, transformação e carga de dados. Tal ferramenta permite ao usuário preparar, limpar, misturar diversos dados de qualquer fonte, ao mesmo tempo que elimina codificações e a complexidade do trabalho com os dados. A maior parte dos dados utilizados para a carga dos perfis dos pesquisadores foram extraídos diretamente das pastas que continham os arquivos XMLs dos pesquisadores baixados da plataforma. Porém, para alguns dados dentro do XML, foram necessários uma exportação via consulta xQuery e a criação de um novo XML para que a ferramenta ETL

[19] Disponível em: http://www.pentaho.com/product/data-integration

conseguisse realizar a importação. Exemplo: A *tag* XML <AREAS-DE-ATUACAO> que contém a área de atuação dos pesquisadores, mesmo sendo obrigatória segundo as definições, não estava presente em alguns arquivos. Ao realizar a carga diretamente dos XML baixados, a ferramenta ETL relatava erro por não encontrar a *tag*, que deveria estar presente no arquivo mesmo que vazia, ou seja, sem informação. Todo esse processo pode ser melhor visualizado no Apêndice B.

Com os dados já importados para dentro das tabelas, foram calculados os elementos e as categorias que compõem o modelo do perfil quantitativo de cada pesquisador, e já computando sua reputação pelo índice Rep-Index [1], [5], [11], conforme as Equações (15) e (16).

$$Rep - Index_{(R)} = \sum_{i=1}^{c} \left(\sum_{j=1}^{e_i} \frac{(v_j . w_j)}{\max(v_j)} \right) \tag{15}$$

Fonte: Cervi, Galante e Oliveira [1], [5], [11].

Sendo, R a reputação do pesquisador, c representa o número total de categorias, i representa o intervalo de 1 até o total de categorias c, e_i representa o número total de elementos em cada categoria, j é o intervalo de 1 até o número total de elementos e_i, v refere-se ao valor do elemento, w_j é o valor do peso do elemento, $max(v_j)$ representa o maior valor do elemento.

A Equação (16) classifica a reputação dos pesquisadores através de cinco níveis, sendo nível 1 para pesquisadores iniciantes e nível 5 para pesquisadores experientes.

$$Rep - Index_{(R)} = \begin{cases} 1, se\ Rep - Index_{(R)} \geq 0\ ^\wedge\ < 20 \\ 2, se\ Rep - Index_{(R)} \geq 20\ ^\wedge\ < 40 \\ 3, se\ Rep - Index_{(R)} \geq 40\ ^\wedge\ < 60 \\ 4, se\ Rep - Index_{(R)} \geq 60\ ^\wedge\ < 80 \\ 5, se\ Rep - Index_{(R)} \geq 80\ ^\wedge\ \leq 100 \end{cases} \tag{16}$$

Fonte: Cervi, Galante e Oliveira [1], [5], [11].

Para o cálculo dos elementos utilizou-se os pesos para o modelo de perfil determinados no estudo proposto por Vivian, Cervi e Rovadosky [47]. Neste trabalho utilizou- se de técnicas de mineração de dados e aprendizado de máquinas presentes no

software *Weka*[20] para computar as opções de peso, buscando assim, os melhores pesos para todos os elementos do modelo de perfil. Foram utilizados os pesos propostos com a técnica *ReliefF*, onde este algoritmo apresenta como característica não descartar atributos, dessa forma, todos os elementos do modelo de perfil foram utilizados para determinar as similaridades entre os pesquisadores.

Um dos elementos da categoria de publicação refere-se à rede de coautoria do pesquisador, ou seja, representa o número de pesquisadores que possuem trabalhos publicados em conjunto com o pesquisador em questão. Esse elemento foi montado da seguinte forma: através de um script leu-se os dados das tabelas de autores de livros, capítulos de livros, artigos em periódicos e trabalhos em eventos, obtendo assim a coautoria do pesquisador. Utilizou-se para desambiguação de nomes, seja por erro de digitação ou outro problema que cause duplicidade de dados, o algoritmo de Levenshtein [48], calibrando o mesmo para considerar equivalente se ambos forem similares em 90%. O mesmo algoritmo foi utilizado para compor os dados textuais dos elementos do perfil.

Posteriormente, baseado nas métricas, calculou-se as similaridades e, por fim, as respectivas recomendações. Para a gravação das similaridades dentro da base de dados estabeleceu-se um limiar, sendo guardadas as comparações que atingissem um coeficiente de correlação forte ou moderado. Para a similaridade do perfil qualitativo que calcula a frequência dos termos e a relevância dos mesmos, utilizou-se das palavras-chave dos elementos do perfil, de cada produção do pesquisador.

3.6. DESENVOLVIMENTO DO SISTEMA WEB SIM(CV)

A implementação da ferramenta ocorreu em etapas. Na primeira etapa, fez-se a importação dos dados e toda a parte de manipulação dos dados do perfil, sendo que na segunda etapa, implementou-se o algoritmo de busca e cálculo de similaridade dos perfis. Na terceira etapa, por sua vez, o algoritmo de recomendação ao usuário gerou as listas de recomendações. E, por fim, na quarta etapa, foi desenvolvido o *front-end* da aplicação. Os principais componentes utilizados para promover o funcionamento da aplicação foram: servidor Web juntamente com linguagem PHP e um banco de dados relacional PostgreSQL.

[20] Disponível em: https://www.cs.waikato.ac.nz/ml/weka/

Para uma melhor experiência do usuário na utilização da interface web, optou-se pelo Bootstrap[21], um *framework* HTML, CSS e JS para desenvolvimento de projetos responsivos.

3.6.1. Visão Geral da Ferramenta

Depois de uma carga inicial massiva de dados, de onde foram descarregados os arquivos XMLs dos pesquisadores e realizado o cálculo de similaridade entre todos, a ferramenta consegue se atualizar conforme demonstrado na Figura 10.

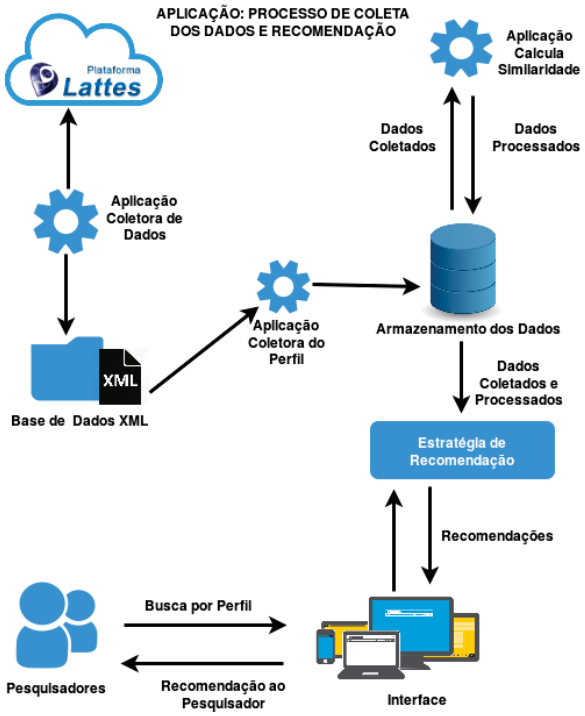


Figura 10: Visão da Aplicação.

Desta forma, existe uma aplicação que coleta os dados da plataforma Lattes e armazena os arquivos dentro de uma pasta, a qual é lida pela aplicação coletora dos dados do perfil, ou seja, verifica os dados do modelo de perfil e guarda na base de dados as informações. Outra aplicação calcula a similaridade dos perfis e armazena na base de dados. Então, é aplicado por outro algoritmo a estratégia de recomendação e, através da interface, o usuário solicita uma busca e recebe as recomendações.

[21] O Bootstrap torna o desenvolvimento *front-end* web mais fácil e rápido, adaptando-se a qualquer dispositivo e tamanho de tela. Disponível em: http://getbootstrap.com.br

3.6.2. Ferramenta SIM(CV)

A tela inicial apresentada na Figura 11 exibe o formulário de pesquisa, através deste busca-se pelo nome do pesquisador para quem se deseja visualizar as recomendações. No menu superior, é possível visualizar um link de informações, o qual detalha melhor as informações sobre as recomendações, relaciona de onde vem os dados e a data de atualização dos mesmos. Ainda no menu superior, foi acrescentado um contador de buscas realizadas dentro do sistema, para uma pequena verificação da utilização da ferramenta.

Figura 11: A tela inicial da aplicação.

Além de informar o nome do pesquisador, informa-se qual tipo de recomendação, a quantidade de recomendações e o limiar desejado de similaridade. No menu inferior, estão disponíveis uma lista das tabelas do sistema. Cada tabela lista os trechos do arquivo DTD, os quais foram utilizados para buscar os dados dentro dos XMLs coletados da plataforma Lattes. Descreve também, todos os campos das tabelas e referencia o seu endereço dentro do arquivo XML, que contém os dados. No outro link, apresenta-se de forma visual o relacionamento das tabelas do sistema.

A Figura 12 apresenta o resultado de uma busca na ferramenta. Na parte superior, pode-se refazer a busca por outro pesquisador. Após o título de “Recomendações para o pesquisador”, pode ser selecionado outro tipo de recomendação para o mesmo perfil do pesquisador. Na parte da lateral esquerda, em verde, estão as informações do perfil do pesquisador buscado. Assim, traz-se o nome, a instituição, a formação, a área cadastrada no Lattes e o cálculo de reputação do pesquisador, pelo uso do Rep-Index. Para maior informação do perfil, também há um link que ao ser clicado aponta para o Lattes do

pesquisador, colocando o cursor sobre o mesmo, aparece o resumo informado no Lattes. Abaixo, visualiza-se a data de atualização do arquivo XML vindo da plataforma Lattes, informações buscadas dentro dos seguintes endereços: /CURRICULO-VITAE/@DATA-ATUALIZACAO e /CURRICULO-VITAE/@HORA-ATUALIZACAO. Na sequência, são apresentados em ordem decrescente os termos ligados ao perfil, já com o cálculo do grau de relevância de cada termo e a frequência em que o termo foi encontrado dentro dos elementos do modelo de perfil.

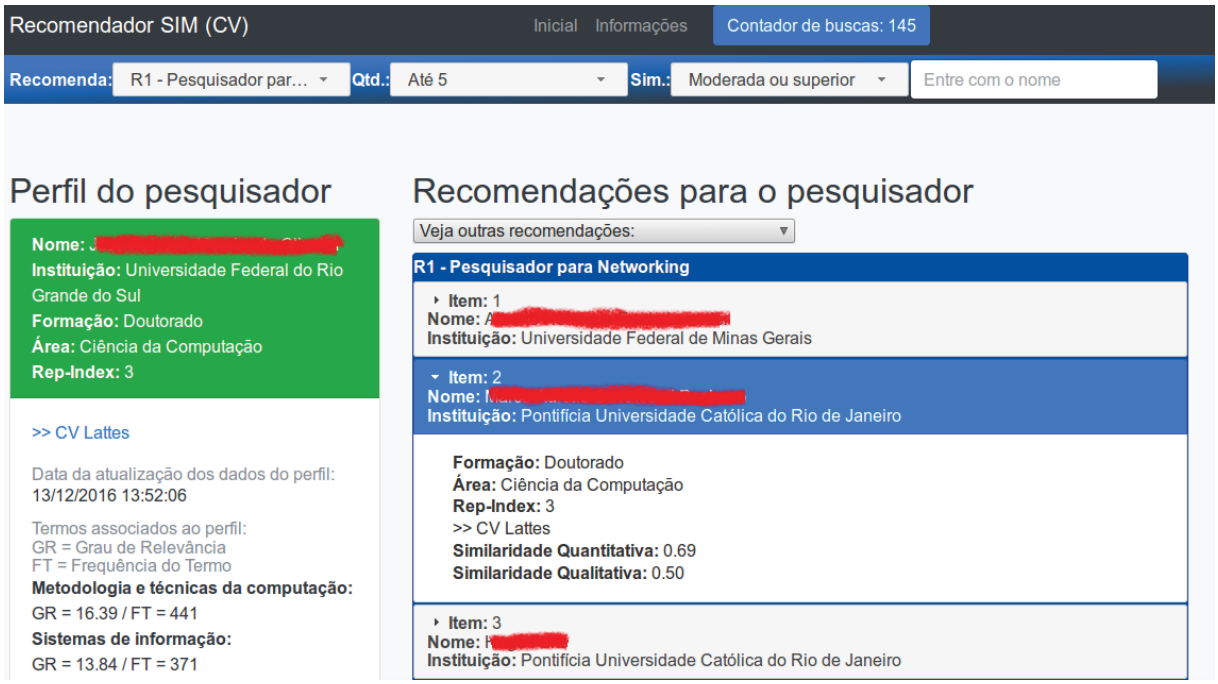


Figura 12: Recomendações para o perfil informado.

Os resultados das recomendações podem ser visualizados na parte direita da tela em azul. Cada item é uma recomendação do tipo de recomendação solicitada. O item em azul mais claro, estará selecionado e mostra mais informação da recomendação como, por exemplo, o Rep-Index, link para o Lattes, e os índice de similaridade quantitativa e qualitativa.

4. EXPERIMENTOS E RESULTADOS

O presente capítulo apresenta os experimentos realizados com o intuito de avaliar a abordagem de recomendação proposta. Para cada experimento é apresentada a metodologia de como foi realizado, seguido de seus resultados. Estes são apresentados em forma de tabela e/ou gráficos e, na sequência, é apresentada análise sobre os resultados obtidos.

Para a avaliação da abordagem, realizou-se os seguintes experimentos: (i) Verificar a capacidade da abordagem de gerar recomendações através do experimento 4.1; (ii) Analisar a capacidade da abordagem de gerar recomendações relevantes através do experimento 4.2; (iii) Analisar se o grau de relevância dos termos acompanhou as mudanças nos interesses do pesquisador através do experimento 4.3.

Ressalta-se que o objetivo dos experimentos é avaliar a abordagem proposta no Capítulo 3. Dessa forma, para todos os experimentos foi utilizado o conjunto de dados obtidos e descritos na Seção 3.5 deste trabalho.

4.1. EXPERIMENTO 1 - CALCULAR A COBERTURA DAS RECOMENDAÇÕES DO SISTEMA

Esse experimento tem como objetivo verificar a capacidade da abordagem de gerar recomendações. A cobertura (*coverage*) trata-se de uma métrica de avaliação muito utilizada em sistemas de recomendação, sendo um dos índices capazes de dizer se a abordagem tem competência para realizar uma boa cobertura de recomendações. Desta forma, a cobertura é a proporção de itens que são aptos de serem recomendados em relação ao conjunto de todos os itens conhecidos pelo sistema de recomendação [49].

Para o cálculo da cobertura foi utilizada a Equação (17), definida por Ge, Delgado-Battenfeld e Jannach [50].

$$cobertura = \frac{|I_p|}{|I|} \tag{17}$$

Fonte: Ge, Delgado-Battenfeld e Jannach [50].

Sendo I o conjunto de todos os itens possíveis do sistema e I_p o conjunto de todos os itens os quais o sistema pode gerar recomendações.

Para realizar o experimento foram utilizadas as recomendações processadas para a ferramenta descrita na Seção 3.6, a qual utilizou um total de 157.601 Currículos Lattes dos pesquisadores das áreas selecionadas. Na Figura 13 visualiza-se o número de pesquisadores por área.

Figura 13: Total de pesquisadores por áreas.

Os resultados da aplicação do cálculo de cobertura podem ser visualizados na Tabela 16, a qual demonstra os valores de acordo com a área do pesquisador, conforme especificado na Seção 3.5 e pelo tipo de recomendação gerada.

Tabela 16. Valores da métrica de cobertura por área para cada abordagem de recomendação.

Área	**R1**	**R2**			**R3**		**R4**		**R5**	**R6**
		A	**B**	**C**	**A**	**B**	**A**	**B**		
Ciência da Computação	0,2757	0,8017	0,7656	0,6886	0,7953	0,7854	0,8673	0,8720	0,2014	0,4578
Genética	0,2396	0,7503	0,7408	0,6693	0,7356	0,7361	0,8025	0,8053	0,1898	0,5620
Engenharia Elétrica	0,2343	0,8200	0,7876	0,6746	0,8022	0,7869	0,8771	0,8684	0,1708	0,2698
Odontologia	0,2472	0,7641	0,7295	0,6259	0,7796	0,7643	0,8157	0,8091	0,2199	0,6237
Agronomia	0,2448	0,8700	0,8666	0,8418	0,8630	0,8636	0,9095	0,9395	0,2245	0,5203
Economia	0,2042	0,8133	0,7689	0,5400	0,8099	0,7960	0,8738	0,8166	0,1034	0,3588
Educação	0,2259	0,8314	0,8122	0,7635	0,8131	0,8112	0,9131	0,9357	0,1087	0,4710
Letras	0,3154	0,8392	0,8215	0,8380	0,8386	0,8371	0,9050	0,9379	0,1974	0,5854
MÉDIA	**0,2483**	**0,8112**	**0,7865**	**0,7052**	**0,8046**	**0,7975**	**0,8705**	**0,8730**	**0,1769**	**0,4811**

Na Tabela 16 as recomendações R2, R3 e R4 foram divididas em A, B e C para um melhor entendimento, pois as mesmas tem recomendações por nível de formação do pesquisador ficando, então, desta forma:

- R1: recomenda para o pesquisador outro pesquisador para networking;
- R2 (A): recomenda para o pesquisador outro pesquisador para parceria em orientações de mestrado;
- R2 (B): recomenda para o pesquisador outro pesquisador para parceria em orientações de doutorado;
- R2 (C): recomenda para o pesquisador outro pesquisador para parceria em orientações de pós-doutorado;
- R3 (A): recomenda para o pesquisador outro pesquisador para participação em bancas de mestrado;
- R3 (B): recomenda para o pesquisador outro pesquisador para participação em bancas de doutorado;
- R4 (A): recomenda para o pesquisador outro pesquisador para orientação em uma futura formação de doutorado;
- R4 (B): recomenda para o pesquisador outro pesquisador para orientação em uma futura formação de pós-doutorado;
- R5: recomenda para o pesquisador a produção dos últimos 5 anos, de outro pesquisador;
- R6: recomenda periódicos em que outros pesquisadores com perfis semelhantes costumam publicar.

Na Figura 14 visualiza-se a métrica de cobertura por recomendações em cada área.

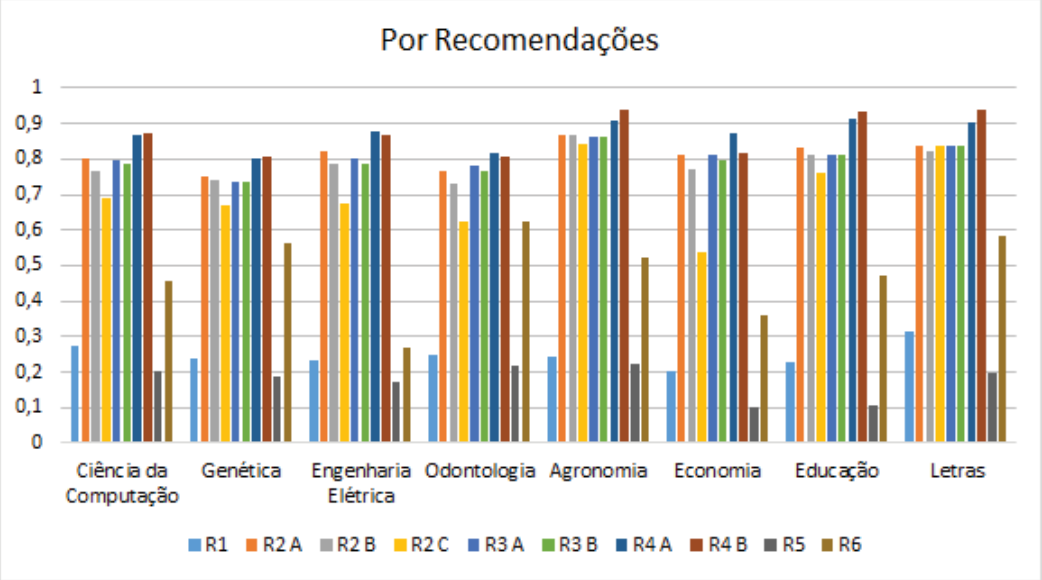


Figura 14: Valores da métrica de cobertura por recomendações em cada área.

Observa-se na Figura 14 uma semelhança no gráfico entre as áreas envolvidas, demonstrando que para a cobertura das recomendações, os elementos do modelo de perfil estão bem ajustados e, desta forma, seus pesos não interferem nas peculiaridades que cada área possui. Na Figura 15 visualiza-se a métrica de cobertura por áreas em cada recomendação.

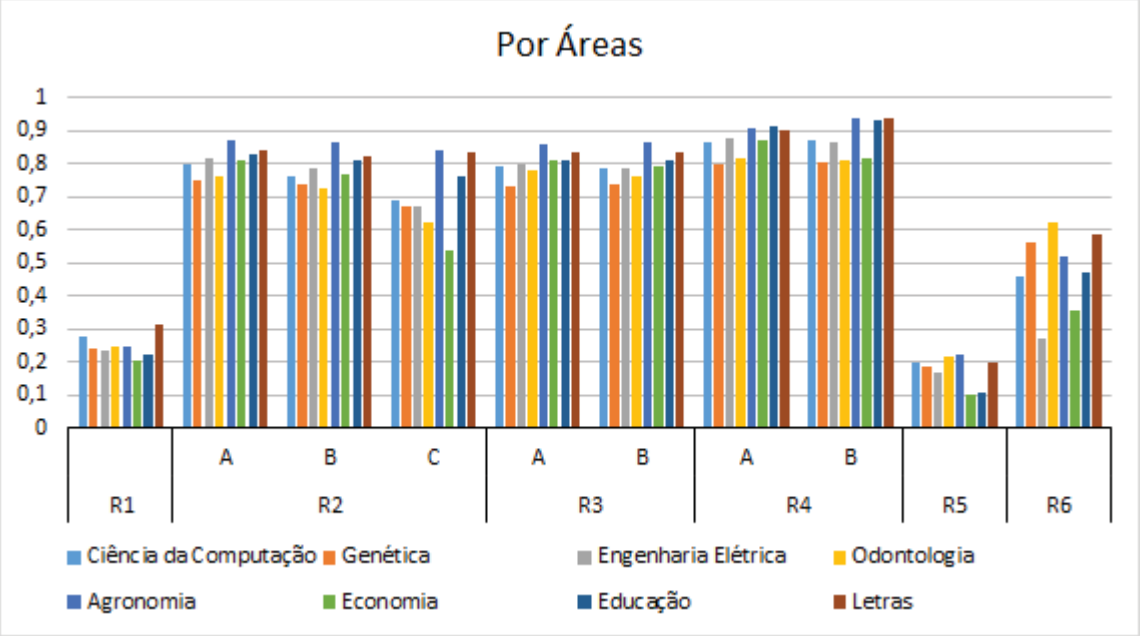


Figura 15: Valores da métrica de cobertura por áreas em cada recomendação.

Na Figura 15 fica evidenciado que a cobertura das recomendações R2, R3, R4 e R6 são satisfatórias. Porém, a cobertura das recomendações R1 e R5 ficaram abaixo das

demais recomendações, em todas as áreas. Pesquisando o ocorrido nessas recomendações, verificou-se que a amplitude dos itens possíveis para recomendação impactam no resultado final. Isso ocorre porque 15% de todos os currículos processados somente continham um elemento do modelo de perfil: o elemento formação acadêmica. Desta forma, a similaridade permanece no patamar de fraca ou ausente, o que é insuficiente para gerar a recomendação. Já as recomendações R2, R3, R4 e R6, por se tratarem de recomendações atreladas a elementos do modelo de perfil para sua realização, esses 15% não entram no cálculo dos itens possíveis do sistema. Isso porque os 15% dos currículos que não possuem tais dados dos elementos não fazem parte do conjunto de itens de possíveis recomendações. O problema da falta de dados para a geração de recomendações em sistemas de recomendação é análogo a todos os sistemas. Para esses 15% podem ser geradas recomendações ditas não personalizadas, como por exemplo, recomendar os perfis mais recomendados do sistema, ou os perfis dos pesquisadores com maior reputação (Rep-Index) da área.

Para verificar se os 15% dos currículos com poucos dados interferiam no cálculo da cobertura da abordagem, realizou-se novamente o processo de cálculo utilizando somente um grupo específico que já estava contido no sistema. Tal grupo deveria ter elementos em todas as categoria. Assim, optou-se por calcular a cobertura dos bolsistas de produtividade da Ciência da Computação. Os dados de quais currículos constituíam os bolsistas de produtividade desta área não estavam presentes na base original. Portanto, precisou-se correlacionar os identificadores únicos dos pesquisadores com a lista dos bolsistas de produtividades da área da Ciência da Computação. Tal processo de correlação está descrito no Apêndice C.

Para o segundo cálculo da cobertura da abordagem utilizou-se 451 currículos de pesquisadores já existentes na base. Estes, por sua vez, foram correlacionados com a lista de bolsistas de produtividade da Ciência da Computação da Plataforma Lattes. A Tabela 17 mostra os valores obtidos da métrica de cobertura deste grupo de pesquisadores.

Tabela 17. Valores da métrica de cobertura para cada abordagem de recomendação dos bolsistas de produtividade da Ciência da Computação.

Área	**R1**	**R2**			**R3**		**R4**		**R5**	**R6**
		A	**B**	**C**	**A**	**B**	**A**	**B**		
Bolsista Produtividade - Ciência da Computação	0,7339	0,6718	0,6555	0,6397	0,6562	0,6535	0	0,7719	0,7244	0,8760

Cabe ressaltar na Tabela 17 a recomendação R4 (A). Esta recomendação para este grupo de pesquisadores apresentou cobertura zero, ou seja, não concretizou nenhuma recomendação. Isso ocorreu, devido ao fato dessa recomendação referir-se à indicação de orientador para uma possível formação futura de doutorado. Porém, neste grupo selecionado, todos já são doutores. A Figura 16 visualiza melhor o resultado com esse segmento de itens para o cálculo da cobertura do sistema.

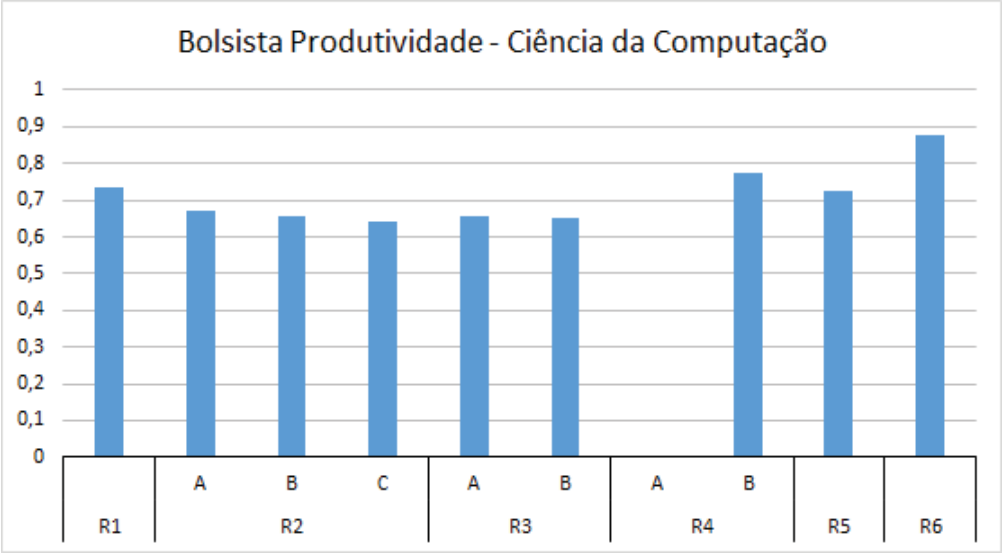


Figura 16: Valores da métrica de cobertura para bolsistas de produtividade da Ciência da Computação

A Figura 16 deixa claro que as recomendações R1 e R5 que obtiveram pouca cobertura observado na Figura 15, agora tem uma cobertura satisfatória como as demais recomendações.

4.2. EXPERIMENTO 2 – CALCULAR A PRECISÃO, REVOCAÇÃO E *F-MEASURE* DO SISTEMA

Esse experimento objetiva verificar a capacidade do sistema de gerar recomendações relevantes para o usuário, através do cálculo das seguintes métricas: (i) precisão (*precision*) que verifica a quantidade de itens recomendados que são do interesse do usuário em relação ao conjunto de todos os itens que lhes são recomendados; (ii) revocação (*recall*) que indica a quantidade de itens de interesse do usuário que aparecem na lista de relevantes; (iii) *F-measure* que é a média harmônica entre a precisão e a revocação e, desta forma, combina-se os dois valores em um para simplificar possíveis comparações.

Para o cálculo da precisão foi utilizada a Equação (18), definida por Huang *et al.* [51].

$$precisão = \frac{\left|R_g \cap R_r\right|}{R_r} \tag{18}$$

Fonte: Huang *et al.* [51].

Para o cálculo da revocação foi utilizada a Equação (19), definida por Huang *et al.* [51].

$$revocação = \frac{\left|R_g \cap R_r\right|}{R_g} \tag{19}$$

Fonte: Huang *et al.* [51].

Sendo R_g o conjunto total relevante, R_r o conjunto de recomendações do sistema e $R_g \cap R_r$ é conjunto de recomendações corretas ou relevantes.

Para o cálculo da *F-measure* foi utilizada a Equação (20), definida por Huang *et al.* [51].

$$F - measure = 2\left(\frac{precisão\ .revocação}{precisão + revocação}\right) \tag{20}$$

Fonte: Huang *et al.* [51].

Para os cálculos das métricas, o experimento foi realizado de forma *offline*, ou seja, sem interação com os usuários. Para isso, utilizou-se o conjunto de dados coletados e gerados previamente. Então, empregou-se os dados a respeito de uma seleção de itens, que descrevem as preferências dos pesquisadores em determinado período de tempo e, comparou-se com o comportamento desses pesquisadores num período de tempo posterior. Assim, avaliando suas preferências posteriores, pode-se prever se a recomendação gerada com base no primeiro período de tempo é de interesse do pesquisador. Esse tipo de experimento tem a vantagem de não necessitar de interação com usuários reais, permitindo que se comparem diversos algoritmos de forma a acelerar o processo de comparação. Ainda, há a vantagem da utilização de um grande número de usuários, o que não se conseguiria de outra forma. Porém, os experimentos *offline* permitem somente a avaliação de uma fração das características de um sistema como, por exemplo, a predição dos algoritmos, não considerando a influência do sistema no comportamento dos usuários. Esta avaliação é tida como inicial para a avaliação

do algoritmo e aqui sugere-se para um trabalho futuro um estudo com usuários reais, ou seja, de forma *online*, onde o usuários avalia cada recomendação.

Para a realização do experimento, os dados dos pesquisadores foram coletados conforme metodologia apresentada na Seção 3.5. Porém, utilizou-se somente os dados dos pesquisadores bolsistas de produtividade da área da Ciência da Computação, correlacionados conforme Apêndice C totalizando, assim, 451 pesquisadores para este experimento. Desta forma, dividiu-se em dois intervalos de tempo os dados.

O primeiro conjunto compreende todo o currículo do pesquisador até o ano de 2012, sendo esse intervalo de tempo utilizado para a geração das recomendações. Ressalta-se que todo o processo descrito na Figura 9, nos passos 9 e 10, tiveram que ser reexecutados levando em consideração até o ano de 2012. Assim, recalculou-se também todos os elementos e categorias do modelo de perfil, bem como todo grau de relevância dos termos dos perfis, também recalculando a similaridade quantitativa e qualitativa, gerando novas recomendações.

No segundo intervalo de tempo, compreendido entre 2013 a 2017, os dados foram utilizados para verificar se o que foi recomendado no primeiro intervalo realmente se confirmou no currículo do pesquisador. Esse segundo período é denominado de conjunto verdade, pois o mesmo é usado para a avaliação da qualidade das recomendações produzidas. Os intervalos de tempo foram escolhidos com o objetivo de analisar se o algoritmo de recomendação consegue predizer as ações que ocorreram no outro intervalo de tempo. Desta forma, utilizando apenas o conjunto base, verifica-se se é possível prever se as recomendações ocorreram no conjunto verdade.

No intervalo de tempo onde foram verificadas as recomendações, chamado de conjunto verdade, são observadas as seguintes questões ligadas às recomendações geradas.

Recomendação 1: recomendação de outro pesquisador. Verificou-se a ocorrência de coautoria nas produções em conjunto entre os pesquisadores. A coautoria foi verificada pelo número de identificador único dos autores nesta recomendação, pois verificou-se que para essa parcela de pesquisadores em anos superiores a 2012, 95% dos autores continham este dado. Observou-se os itens abaixo listado no currículo do pesquisador em artigos em periódicos, capítulos de livros, livros, trabalhos em eventos, projetos de pesquisas e produção de software, cruzando-se com as recomendações realizadas ao pesquisador alvo. As localizações dos dados extraídos dos XMLs são listadas a seguir, no formato XPath:

- Identificador único do pesquisador alvo da recomendação:
  - ✓ /CURRICULO-VITAE/@NUMERO-IDENTIFICADOR

- Ano da publicação > 2012:
  - ✓ /CURRICULO-VITAE/PRODUCAO-BIBLIOGRAFICA/ARTIGOS-PUBLICADOS/ARTIGO-PUBLICADO/DADOS-BASICOS-DO-ARTIGO/@ANO-DO-ARTIGO
  - ✓ /CURRICULO-VITAE/PRODUCAO-BIBLIOGRAFICA/LIVROS-E-CAPITULOS/CAPITULOS-DE-LIVROS-PUBLICADOS/CAPITULO-DE-LIVRO-PUBLICADO/DADOS-BASICOS-DO-CAPITULO/@ANO
  - ✓ /CURRICULO-VITAE/PRODUCAO-BIBLIOGRAFICA/LIVROS-E-CAPITULOS/LIVROS-PUBLICADOS-OU-ORGANIZADOS/LIVRO-PUBLICADO-OU-ORGANIZADO/DADOS-BASICOS-DO-LIVRO/@ANO
  - ✓ /CURRICULO-VITAE/PRODUCAO-BIBLIOGRAFICA/TRABALHOS-EM-EVENTOS/TRABALHO-EM-EVENTOS/DADOS-BASICOS-DO-TRABALHO/@ANO-DO-TRABALHO
  - ✓ /CURRICULO-VITAE/DADOS-GERAIS/ATUACOES-PROFISSIONAIS/ATUACAO-PROFISSIONAL/ATIVIDADES-DE-PARTICIPACAO-EM-PROJETO/PARTICIPACAO-EM-PROJETO/PROJETO-DE-PESQUISA/@ANO-INICIO
  - ✓ /CURRICULO-VITAE/PRODUCAO-TECNICA/SOFTWARE/DADOS-BASICOS-DO-SOFTWARE/@ANO
- Identificador único dos pesquisadores autores das publicações possíveis recomendados para o pesquisador alvo:
  - ✓ /CURRICULO-VITAE/PRODUCAO-BIBLIOGRAFICA/ARTIGOS-PUBLICADOS/ARTIGO-PUBLICADO/AUTORES/@NRO-ID-CNPQ
  - ✓ /CURRICULO-VITAE/PRODUCAO-BIBLIOGRAFICA/LIVROS-E-CAPITULOS/CAPITULOS-DE-LIVROS-PUBLICADOS/CAPITULO-DE-LIVRO-PUBLICADO/AUTORES/@NRO-ID-CNPQ
  - ✓ /CURRICULO-VITAE/PRODUCAO-BIBLIOGRAFICA/LIVROS-E-CAPITULOS/LIVROS-PUBLICADOS-OU-ORGANIZADOS/LIVRO-PUBLICADO-OU-ORGANIZADO/AUTORES/@NRO-ID-CNPQ
  - ✓ /CURRICULO-VITAE/PRODUCAO-BIBLIOGRAFICA/TRABALHOS-EM-EVENTOS/TRABALHO-EM-EVENTOS/AUTORES/@NRO-ID-CNPQ
  - ✓ /CURRICULO-VITAE/DADOS-GERAIS/ATUACOES-PROFISSIONAIS/ATUACAO-PROFISSIONAL/ATIVIDADES-DE-PARTICIPACAO-EM-PROJETO/PARTICIPACAO-EM-PROJETO/PROJETO-DE-PESQUISA/EQUIPE-DO-PROJETO/INTEGRANTES-DO-PROJETO/@NRO-ID-CNPQ
  - ✓ /CURRICULO-VITAE/PRODUCAO-TECNICA/SOFTWARE/AUTORES/@NRO-ID-CNPQ

Recomendação 2: recomendação de pesquisadores para orientação em conjunto. Verificou-se a ocorrência de orientações em mestrado, doutorado ou pós-doutorado em conjunto entre os pesquisadores, cruzando os nomes dos orientados para essa verificação. Foi utilizada a função *similarity* da extensão *pg_trgm* no PostgreSql para localizar nomes até 90% similares, para o caso de os pesquisadores terem informado de forma diferente os nomes de seus orientados. Desta forma, ao se encontrar orientados com pesquisadores diferentes, verificou-se o tipo de orientação, para saber quem era o orientador principal e o co-orientador. As localizações dos dados extraídos dos XMLs são listadas a seguir, no formato XPath:

- Identificador único do pesquisador alvo da recomendação:

- ✓ /CURRICULO-VITAE/@NUMERO-IDENTIFICADOR

- Ano da orientação > 2012:
  - ✓ /CURRICULO-VITAE/OUTRA-PRODUCAO/ORIENTACOES-CONCLUIDAS/ORIENTACOES-CONCLUIDAS-PARA-MESTRADO/DADOS-BASICOS-DE-ORIENTACOES-CONCLUIDAS-PARA-MESTRADO/@ANO
  - ✓ /CURRICULO-VITAE/OUTRA-PRODUCAO/ORIENTACOES-CONCLUIDAS/ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO/DADOS-BASICOS-DE-ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO/@ANO
  - ✓ /CURRICULO-VITAE/OUTRA-PRODUCAO/ORIENTACOES-CONCLUIDAS/ORIENTACOES-CONCLUIDAS-PARA-POS-DOUTORADO/DADOS-BASICOS-DE-ORIENTACOES-CONCLUIDAS-PARA-POS-DOUTORADO/@ANO
- Natureza da orientação:
  - ✓ /CURRICULO-VITAE/OUTRA-PRODUCAO/ORIENTACOES-CONCLUIDAS/ORIENTACOES-CONCLUIDAS-PARA-MESTRADO/DADOS-BASICOS-DE-ORIENTACOES-CONCLUIDAS-PARA-MESTRADO/@NATUREZA
  - ✓ /CURRICULO-VITAE/OUTRA-PRODUCAO/ORIENTACOES-CONCLUIDAS/ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO/DADOS-BASICOS-DE-ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO/@NATUREZA
  - ✓ /CURRICULO-VITAE/OUTRA-PRODUCAO/ORIENTACOES-CONCLUIDAS/ORIENTACOES-CONCLUIDAS-PARA-POS-DOUTORADO/DADOS-BASICOS-DE-ORIENTACOES-CONCLUIDAS-PARA-POS-DOUTORADO/@NATUREZA
- Tipo de orientação:
  - ✓ /CURRICULO-VITAE/OUTRA-PRODUCAO/ORIENTACOES-CONCLUIDAS/ORIENTACOES-CONCLUIDAS-PARA-MESTRADO/DETALHAMENTO-DE-ORIENTACOES-CONCLUIDAS-PARA-MESTRADO/@TIPO-DE-ORIENTACAO
  - ✓ /CURRICULO-VITAE/OUTRA-PRODUCAO/ORIENTACOES-CONCLUIDAS/ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO/DETALHAMENTO-DE-ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO/@TIPO-DE-ORIENTACAO
  - ✓ /CURRICULO-VITAE/OUTRA-PRODUCAO/ORIENTACOES-CONCLUIDAS/ORIENTACOES-CONCLUIDAS-PARA-POS-DOUTORADO/DETALHAMENTO-DE-ORIENTACOES-CONCLUIDAS-PARA-POS-DOUTORADO/@TIPO-DE-ORIENTACAO
- Nome do orientado com 90% de similaridade:
  - ✓ /CURRICULO-VITAE/OUTRA-PRODUCAO/ORIENTACOES-CONCLUIDAS/ORIENTACOES-CONCLUIDAS-PARA-MESTRADO/DETALHAMENTO-DE-ORIENTACOES-CONCLUIDAS-PARA-MESTRADO/@NOME-DO-ORIENTADO
  - ✓ /CURRICULO-VITAE/OUTRA-PRODUCAO/ORIENTACOES-CONCLUIDAS/ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO/DETALHAMENTO-DE-ORIENTACOES-CONCLUIDAS-PARA-DOUTORADO/@NOME-DO-ORIENTADO
  - ✓ /CURRICULO-VITAE/OUTRA-PRODUCAO/ORIENTACOES-CONCLUIDAS/ORIENTACOES-CONCLUIDAS-PARA-POS-DOUTORADO/DETALHAMENTO-DE-ORIENTACOES-CONCLUIDAS-PARA-POS-DOUTORADO/@NOME-DO-ORIENTADO

Recomendação 3: recomendação de membros bancas de avaliação. Verificou-se a ocorrência de participação em bancas de mestrado ou doutorado com os pesquisadores recomendados. Nesse caso, a recomendação foi verificada pelo número de identificador único dos participantes das bancas e também pelo nome do candidato da banca com a função de

similarity do PostgreSql, já que se verificou que para essa parcela de pesquisadores, em anos superiores a 2012, apenas 76% dos participante continham o número identificador informado. As localizações dos dados extraídos dos XMLs são listadas a seguir, no formato XPath:

- Identificador único do pesquisador alvo da recomendação:
  - ✓ /CURRICULO-VITAE/@NUMERO-IDENTIFICADOR
- Ano da participação na banca > 2012:
  - ✓ /CURRICULO-VITAE/DADOS-COMPLEMENTARES/PARTICIPACAO-EM-BANCA-TRABALHOS-CONCLUSAO/PARTICIPACAO-EM-BANCA-DE-MESTRADO/DADOS-BASICOS-DA-PARTICIPACAO-EM-BANCA-DE-MESTRADO/@ANO
  - ✓ /CURRICULO-VITAE/DADOS-COMPLEMENTARES/PARTICIPACAO-EM-BANCA-TRABALHOS-CONCLUSAO/PARTICIPACAO-EM-BANCA-DE-DOUTORADO/DADOS-BASICOS-DA-PARTICIPACAO-EM-BANCA-DE-DOUTORADO/@ANO
- Natureza da banca:
  - ✓ /CURRICULO-VITAE/DADOS-COMPLEMENTARES/PARTICIPACAO-EM-BANCA-TRABALHOS-CONCLUSAO/PARTICIPACAO-EM-BANCA-DE-MESTRADO/DADOS-BASICOS-DA-PARTICIPACAO-EM-BANCA-DE-MESTRADO/@NATUREZA
  - ✓ /CURRICULO-VITAE/DADOS-COMPLEMENTARES/PARTICIPACAO-EM-BANCA-TRABALHOS-CONCLUSAO/PARTICIPACAO-EM-BANCA-DE-DOUTORADO/DADOS-BASICOS-DA-PARTICIPACAO-EM-BANCA-DE-DOUTORADO/@NATUREZA
- Nome do candidato com 90% de similaridade:
  - ✓ /CURRICULO-VITAE/DADOS-COMPLEMENTARES/PARTICIPACAO-EM-BANCA-TRABALHOS-CONCLUSAO/PARTICIPACAO-EM-BANCA-DE-MESTRADO/DETALHAMENTO-DA-PARTICIPACAO-EM-BANCA-DE-MESTRADO/@NOME-DO-CANDIDATO
  - ✓ /CURRICULO-VITAE/DADOS-COMPLEMENTARES/PARTICIPACAO-EM-BANCA-TRABALHOS-CONCLUSAO/PARTICIPACAO-EM-BANCA-DE-DOUTORADO/DETALHAMENTO-DA-PARTICIPACAO-EM-BANCA-DE-DOUTORADO/@NOME-DO-CANDIDATO
- Identificador único dos pesquisadores participantes das bancas possíveis recomendados para o pesquisador alvo:
  - ✓ /CURRICULO-VITAE/DADOS-COMPLEMENTARES/PARTICIPACAO-EM-BANCA-TRABALHOS-CONCLUSAO/PARTICIPACAO-EM-BANCA-DE-MESTRADO/PARTICIPANTE-BANCA/@NRO-ID-CNPQ
  - ✓ /CURRICULO-VITAE/DADOS-COMPLEMENTARES/PARTICIPACAO-EM-BANCA-TRABALHOS-CONCLUSAO/PARTICIPACAO-EM-BANCA-DE-DOUTORADO/PARTICIPANTE-BANCA/@NRO-ID-CNPQ

Recomendação 4: recomendação de orientador para futura formação do pesquisador. Analisou-se a ocorrência de nova formação do pesquisador e se orientador pertence a lista de recomendados. Nos arquivos XMLs dos currículos há a opção da informação do número identificador único do orientador, porém nessa seleção de pesquisadores com ano de início de sua formação superior a 2012, nenhum registro foi

encontrado. Desta forma, utilizou-se novamente a função de similaridade do PostgreSql, para correlacionar o nome dos orientadores. Ressalta-se que a recomendação R4 (A) não foi avaliada visto que a mesma não gerou recomendações, pois todos os pesquisadores desta seleção já possuem doutorado. As localizações dos dados extraídos dos XMLs são listadas a seguir, no formato XPath:

- Identificador único do pesquisador alvo da recomendação:
  - ✓ /CURRICULO-VITAE/@NUMERO-IDENTIFICADOR
- Ano de início da formação > 2012:
  - ✓ CURRICULO-VITAE/DADOS-GERAIS/FORMACAO-ACADEMICA-TITULACAO/DOUTORADO/@ANO-DE-INICIO
  - ✓ /CURRICULO-VITAE/DADOS-GERAIS/FORMACAO-ACADEMICA-TITULACAO/POS-DOUTORADO/@ANO-DE-INICIO
- Nível da formação:
  - ✓ /CURRICULO-VITAE/DADOS-GERAIS/FORMACAO-ACADEMICA-TITULACAO/DOUTORADO/@NIVEL
  - ✓ /CURRICULO-VITAE/DADOS-GERAIS/FORMACAO-ACADEMICA-TITULACAO/POS-DOUTORADO/@NIVEL
- Nome do orientador com 90% de similaridade:
  - ✓ /CURRICULO-VITAE/DADOS-GERAIS/FORMACAO-ACADEMICA-TITULACAO/DOUTORADO/@NOME-COMPLETO-DO-ORIENTADOR
  - ✓ /CURRICULO-VITAE/DADOS-GERAIS/FORMACAO-ACADEMICA-TITULACAO/POS-DOUTORADO/@NOME-COMPLETO-DO-ORIENTADOR

Recomendação 5: recomendação de produções de outros pesquisadores. Essa recomendação não foi avaliada, pois não conseguiu-se verificar no currículo do pesquisador uma forma de validar como correta ou relevante a recomendação de um artigo ao mesmo.

Recomendação 6: recomendação de periódicos para o pesquisador. Analisou-se a ocorrência de publicações nos periódicos recomendados. Para isso, verificou-se na tabela de recomendação os pesquisadores que receberam recomendações de periódicos e, dentre esses, buscou-se os artigos com ano superior a 2012, com ISSN igual ao ISSN recomendado ao pesquisador. Utilizou-se o ISSN para a correlação, pois nessa seleção de pesquisadores, com artigos publicados em anos superiores a 2012, 91% continham o ISSN. As localizações dos dados extraídos dos XMLs são listadas a seguir, no formato XPath:

- Identificador único do pesquisador alvo da recomendação:
  - ✓ /CURRICULO-VITAE/@NUMERO-IDENTIFICADOR
- Ano do artigo > 2012:
  - ✓ /CURRICULO-VITAE/PRODUCAO-BIBLIOGRAFICA/ARTIGOS-PUBLICADOS/ARTIGO-PUBLICADO/DADOS-BASICOS-DO-ARTIGO/@ANO-DO-ARTIGO

- ISSN do periódico em que o artigo foi publicado:
  - ✓ /CURRICULO-VITAE/PRODUCAO-BIBLIOGRAFICA/ARTIGOS-PUBLICADOS/ARTIGO-PUBLICADO/DETALHAMENTO-DO-ARTIGO/@ISSN

Os resultados das métricas utilizadas neste experimento podem ser observados na Tabela 18, na qual são apresentados os valores das métricas de avaliação por recomendação.

Tabela 18. Valores das métricas de avaliação para os bolsistas de produtividade da área de Ciência da Computação.

Recomendação	**Precisão**	**Revocação**	***F-Measure***
R1	0,0472	0,0344	0,0396
R2 (A)	0,0304	0,3870	0,0562
R2 (B)	0,0197	0,3058	0,0370
R2 (C)	0,0126	0,3636	0,0240
R3 (A)	0,0350	0,0639	0,0452
R3 (B)	0,0376	0,0855	0,0522
R4 (A)	-	-	-
R4 (B)	0,0052	0,1698	0,0100
R5	-	-	-
R6	0,0397	0,5202	0,0736

A Tabela 18 apresenta os dados das métricas de precisão e revocação adaptados ao contexto do experimento, sendo que a relação de precisão se deu através das recomendações relevantes dos fatos buscadas no conjunto verdade pelo total de recomendações feitas. Desta forma, a precisão obteve resultado baixo porém, esperados, visto que é difícil mensurar a relevância de uma recomendação somente medindo a ocorrência de acontecimentos entre os perfis dentro dos elementos do currículo. Pode-se citar como exemplo a recomendação R1, a qual recomenda um perfil de pesquisador para *networking*. Tal recomendação pode ser relevante ao pesquisador que a recebeu, mas não necessariamente os mesmos irão produzir em conjunto. Contudo, foi esta a forma encontrada para verificar no conjunto verdade e assumir que essa recomendação foi relevante.

Já a revocação deu-se através das recomendações relevantes dos fatos encontrados no conjunto verdade pelo total dos fatos encontrados no conjunto verdade, assumindo estes como sendo o conjunto total relevante ao pesquisador. Porém, não somente isso pode ser

relevante ao pesquisador, pois itens desconhecidos também podem ser relevantes e, desta forma, há a tendência de os valores ficarem baixos para a revocação.

O método *offline* utilizado permitiu verificar o potencial dos algoritmos de recomendação até certo nível. Porém, o mérito da abordagem só pode ser medido completamente através da avaliação do usuário após as recomendações. Mesmo que os resultados tenham apresentado valores baixos, evidencia-se que os resultados alcançados de forma inicial justificam estudos aplicados junto a usuários reais que possam fornecer *feedback* das recomendações para análises mais profundas. A falta do *feedback* e o comportamento dos usuários frente as recomendações também explicam os valores baixos obtidos.

Ao analisar os fatos ocorrido em duas partes de tempo utilizando dados reais, valores altos de precisão e revocação poderiam evidenciar a não necessidade de um sistema de recomendação, pois as descobertas de informação estariam se comportando de maneira natural, já potencializando a sua própria descoberta de informação. Já valores baixos, porém não zerados, mostram que a abordagem tende a funcionar, aumentando a descoberta de informação útil ao usuário.

4.3. EXPERIMENTO 3 – ANALISAR A RELEVÂNCIA DO TERMO PARA O PERFIL DOS PESQUISADORES CONSIDERANDO ASPECTOS TEMPORAIS

O objetivo desse experimento é verificar se o grau de relevância dos termos identificados no currículo do pesquisador sofreu alterações em diferentes momentos do tempo, através de duas avaliações. A primeira, utilizou-se dos dados dos bolsistas de produtividade da área de Ciência da Computação (Apêndice C) e a segunda, utilizou-se todos os pesquisadores das oito áreas descritas na Seção 3.5.

Espera-se com o experimento que os termos encontrados no perfil em anos recentes tenham grau de relevância maior e, que os termos que não estão ativos no período, porém presentes em anos anteriores, vão regredindo seu valor com o passar dos anos, expressando, assim, menos interesse pelo termo. Para o experimento, buscou-se o pesquisador com maior Rep-Index absoluto, conforme Equação (15), de cada área coletada pelo trabalho.

Para primeira avaliação, definiu-se três períodos de tempo com intervalo de quatro anos entre cada período. A comparação permitiu uma análise temporal de ocorrência de termos na carreira do pesquisador, possibilitando que seja identificado se ocorreram variações no grau relevância do termo. Os intervalos de tempo utilizados foram do ano de início do

currículo do pesquisador até 2009 (Intervalo 1), até 2013 (Intervalo 2) e até 2017 (Intervalo 3). Analisou-se os sete maiores termos do pesquisador nos referidos intervalos de tempo. Os resultados podem ser visualizados na Tabela 19.

Tabela 19. Variação do grau de relevância dos termos em três períodos, do pesquisador com maior rep-index absoluto, dos bolsistas de produtividade da área de Ciência da Computação.

Termos	Intervalo 1 - Até 2009		Intervalo 2 - Até 2013		Intervalo 3 - Até 2017	
	GR	Ranking	GR	Ranking	GR	Ranking
A	4,3840	4	10,8425	1	25,7634	1
B	7,4320	1	9,8188	2	13,3063	2
C	0,0266	163	2,0516	7	9,0891	3
D	6,0616	2	6,2357	3	8,2607	4
E	4,5689	3	4,7726	4	6,9226	5
F	2,1524	6	3,055	5	3,8432	6
G	2,6552	5	2,9215	6	3,6465	7

Na Tabela 19 visualiza-se os três períodos de tempo, para os sete termos com maior grau de relevância achado no intervalor 3, o qual é período mais atual dos termos do pesquisador. Os dados foram colocados em ordem decrescente gerando um ranking de preferência dos termos. Posteriormente, foram buscados nos intervalos de tempos anteriores o valor do grau de relevância destes mesmos termos e a posição que ocupavam no ranking.

Se visualizarmos apenas as linhas das tabelas, verifica-se que os termos aumentam seu valor conforme o tempo vai passando na carreira do pesquisador. Isso acontece porque o pesquisador continuou atualizando em seus elementos do perfil tal termo. Se observarmos as colunas de ranking identificamos que, por mais que todos os termos tenham aumentado seu grau de relevância com o passar do tempo, ainda existe a preferência diferente ou interações maiores com o termo em determinado momento do que com outro com o passar dos anos. Para esse pesquisador, o termo C foi o que mais variou em seus intervalos de tempo no ranking. No primeiro intervalo esse termo ocupava a posição 163, ou seja, haviam 162 termos com maior preferência para o pesquisador. Já no segundo intervalo o termo C já ocupava a posição 7 do ranking, e no intervalo mais atual, a posição 3 do ranking. Verifica-se que esse termo, com o passar do anos, foi muito utilizado pelo pesquisador e o mesmo expressa uma de suas preferências. Isso evidencia que o cálculo utilizado para a verificação do grau de

relevância para o termo consegue acompanhar as mudanças junto ao currículo do pesquisador com o passar do tempo.

Para a segunda avaliação, definiu-se um período de cinco anos no currículo dos pesquisadores envolvidos. O marco inicial do currículo aconteceu em 2013, ou seja, começou-se a verificação dos termo a partir de 2013. Então, calculou-se novamente os termos a partir de 2013 até 2014, até 2015, até 2016 e por fim até 2017. Assim, tem-se um aspecto de comportamento dos últimos cinco anos dos termos. Da mesma, analisou-se os cinco maiores termos do pesquisador no primeiro intervalo de tempo (2013) e acompanhou-se esses termos com o passar dos anos. Os resultados podem ser visualizados nas Tabelas de 20 a 27.

Tabela 20. Variação do grau de relevância dos termos no período de cinco anos do pesquisador com maior rep-index da área de Ciência da Computação.

Termos	2013		Até 2014		Até 2015		Até 2016		Até 2017	
	GR	Rank	GR	Rank	GR	Rank	GR	Rank	GR	Rank
A	1,9300	1	4,8175	1	8,5508	1	13,4508	1	16,8503	1
B	0,7500	2	2,0875	2	3,6875	2	6,9375	2	7,7875	2
C	0,6100	3	1,3225	3	2,0725	3	2,7475	3	4,0975	3
D	0,3800	4	0,7925	4	1,1925	4	2,2675	4	3,1175	4
E	0,2400	5	0,4775	5	0,5275	10	1,2775	7	1,8275	7

A Tabela 20 demonstra que as maiores preferências e interações com os 5 termos com maior grau de relevância para esse pesquisador da área da Ciência da Computação, pouco mudou em 5 anos de seu currículo, mostrando uma estabilidade de suas preferências. O termo E, apesar de continuar sendo utilizado e aumentando seu grau de relevância, ainda assim mudou de posição no ranking de preferências.

Tabela 21. Variação do grau de relevância dos termos no período de cinco anos do pesquisador com maior rep-index da área de Genética.

Termos	2013		Até 2014		Até 2015		Até 2016		Até 2017	
	GR	Rank	GR	Rank	GR	Rank	GR	Rank	GR	Rank
A	1,3600	1	3,4600	1	6,7833	1	10,7458	1	12,6458	1
B	0,7200	2	2,0200	2	3,6500	2	6,4875	2	6,8875	2
C	0,3200	3	0,3200	14	0,3200	27	0,5200	24	0,5200	37
D	0,3200	3	0,6200	3	1,0500	4	2,0875	3	2,4875	5

Termos	2013		Até 2014		Até 2015		Até 2016		Até 2017	
	GR	Rank	GR	Rank	GR	Rank	GR	Rank	GR	Rank
E	0,24	5	0,3400	11	0,4733	12	0,6733	13	1,7233	9

Para o pesquisador da área de Genética avaliado na Tabela 21, verifica-se uma grande variação do ranking de seus termos mais utilizados. Isso deverá impactar em suas similaridades e recomendações com o passar do tempo.

Tabela 22. Variação do grau de relevância dos termos no período de cinco anos do pesquisador com maior rep-index da área de Engenharia Elétrica.

Termos	2013		Até 2014		Até 2015		Até 2016		Até 2017	
	GR	Rank	GR	Rank	GR	Rank	GR	Rank	GR	Rank
A	1,3600	1	3,4600	1	6,7833	1	10,7458	1	12,7958	1
B	0,7200	2	2.0200	2	3,6500	2	6,4875	2	6,8875	5
C	0,6400	3	1,0400	4	1,4400	4	1,6400	6	2,0400	12
D	0,4800	4	0,4800	6	0,4800	14	0,4800	25	1,6800	17
E	0,3500	5	1,0500	3	3.0500	3	4,6000	3	7,7000	4

Na Tabela 22 também se observa uma variação no ranking com o passar do tempo.

Tabela 23. Variação do grau de relevância dos termos no período de cinco anos do pesquisador com maior rep-index da área de Odontologia.

Termos	2013		Até 2014		Até 2015		Até 2016		Até 2017	
	GR	Rank	GR	Rank	GR	Rank	GR	Rank	GR	Rank
A	1,3700	1	3,3075	1	5,1741	1	5,6491	2	7,0491	3
B	1,0200	2	2,7200	2	4,3366	2	5,5366	3	8,1366	2
C	0,9900	3	2,0900	3	3,7066	3	4,5066	5	6,1066	6
D	0,5900	4	1,6025	4	3,5025	4	5,1775	4	6,5275	4
E	0,5100	5	1,3100	6	2,6600	6	3,6600	6	5,9600	7

A Tabela 23, referente ao pesquisador avaliado da área de Odontologia, apresenta uma variação menor em seus termos. Porém, nos últimos dois anos, o primeiro termo do ranking não estava entre os termos mais utilizados pelo pesquisador nos últimos três anos.

Esses movimentos devem ser captados pelo sistema para sugerir conteúdos de acordo com as novas preferências do usuário.

Tabela 24. Variação do grau de relevância dos termos no período de cinco anos do pesquisador com maior rep-index da área de Agronomia.

Termos	2013		Até 2014		Até 2015		Até 2016		Até 2017	
	GR	Rank	GR	Rank	GR	Rank	GR	Rank	GR	Rank
A	0,8700	1	1,7950	2	3,9783	1	6,1533	2	16,0533	1
B	0,4000	2	0,5000	13	0,6333	19	0,6333	26	0,6333	54
C	0,3100	3	0,4225	15	1,8391	8	3,8641	4	6,0641	8
D	0,3000	4	0,3375	18	0,6541	18	0,6541	25	7,5041	5
E	0,2900	5	0,6525	9	1,4691	9	1,4691	10	8,0191	4

O pesquisador avaliado da área de Agronomia, na Tabela 24, também apresentou uma grande variação no ranking dos termos avaliados. O termo B que em 2013 estava em segundo com maior grau de relevância, 4 anos após estava na posição 54.

Tabela 25. Variação do grau de relevância dos termos no período de cinco anos do pesquisador com maior rep-index da área de Economia.

Termos	2013		Até 2014		Até 2015		Até 2016		Até 2017	
	GR	Rank	GR	Rank	GR	Rank	GR	Rank	GR	Rank
A	0,8100	1	1,5850	2	2,5433	4	4,7183	2	4,0133	5
B	0,4000	2	0,5000	12	0,6333	18	0,6333	24	0,6333	37
C	0,3200	3	0,3575	17	0,4075	26	0,4075	36	2,3575	11
D	0,2800	4	0,3925	15	1,3466	9	3,2591	7	3,0266	9
E	0,2400	5	0,5400	10	0,8066	13	0,8066	17	0,8066	31

Na Tabela 25, para o pesquisador da área de Economia a variação entre os termos foi grande o suficiente para que somente o termo A, que estava presente entre os cinco maiores em 2013, permanece-se presente nos anos seguintes.

Tabela 26. Variação do grau de relevância dos termos no período de cinco anos do pesquisador com maior rep-index da área de Educação.

Termos	2013		Até 2014		Até 2015		Até 2016		Até 2017	
	GR	Rank	GR	Rank	GR	Rank	GR	Rank	GR	Rank
A	1,9600	1	3,8475	1	4,3308	2	8,0558	2	37,7558	2
B	1,6400	2	3,315	3	3,7150	3	7,5150	3	39,315	1
C	1,4900	3	3,5400	2	4,5233	1	10,9733	1	24,5233	3
D	0,7100	4	0,8225	7	1,0225	7	1,2975	11	9,3975	4
E	0,6000	5	1,0750	6	1,2583	6	1,6083	8	3,8583	9

A Tabela 26 não apresentou muitas variações na posição do ranking dos termos, mas identificou um grande aumento no grau de relevância dos termos entre 2016 e 2017, mostrando que a produção referente a esses termos para o pesquisador da área de Educação é importante e expressa suas preferências.

Tabela 27. Variação do grau de relevância dos termos no período de cinco anos do pesquisador com maior rep-index da área de Letras.

Termos	2013		Até 2014		Até 2015		Até 2016		Até 2017	
	GR	Rank	GR	Rank	GR	Rank	GR	Rank	GR	Rank
A	0,8500	1	1,8000	4	2,4666	4	3,5666	4	13,1166	2
B	0,7300	2	2,9300	1	4,2633	1	5,8633	1	8,0133	5
C	0,7000	3	2,3250	3	2,8583	3	4,2583	3	14,5583	1
D	0,6500	4	2,7125	2	3,7791	2	5,3791	2	8,2791	3
E	0,5600	5	0,5600	13	0,5600	16	0,5600	21	0,7600	22

A Tabela 27 também mostra uma boa variação do ranking dos termos do pesquisador referente à área de Letras.

Diante dos experimentos, verificou-se que os principais termos dos pesquisadores das 8 áreas mudaram não somente de valor, mas também no ranking de posicionamento dos maiores graus de relevância dos termos. Isso demonstra que o cálculo consegue capturar mudanças nos termos dos elementos do perfil do pesquisador. Se no caso for considerado que os termos que o pesquisador mais utiliza são os termos que ele prefere, é possível capturar as preferências do pesquisador e suas mudanças com o passar do tempo. Tais mudanças devem refletir nos cálculos de similaridade para que o sistema consiga recomendar conteúdos úteis

ao pesquisador e se consiga, com o passar do tempo, continuar com as descobertas sem que o mesmo precise realizar intervenções no sistema, apenas bastando alimentar dados em seu currículo.

5. CONSIDERAÇÕES FINAIS

Nesta seção são apresentadas as considerações finais, destacando-se os objetivos, as contribuições e os resultados obtidos pela abordagem proposta. Por fim, são discutidas algumas sugestões de trabalhos futuros identificados ao longo do desenvolvimento do trabalho.

É importante destacar que com foco nos desafios de recomendar conteúdo útil para pesquisadores que estão sempre se atualizando e se desenvolvendo na carreira, tarefa considerada nada fácil pela quantidade de conteúdo produzido pelos mesmos, o presente trabalho procurou entender e identificar como o perfil de um pesquisador pode se comportar e como entender o mecanismo das preferências que esse perfil exprime sobre o pesquisador. Um dos objetivos principais residiu em identificar uma forma de capturar as mudanças das preferências do pesquisador com o passar dos anos e, de forma mais transparente possível, entender e predizer o que o mesmo pode vir a querer agregar para a atualização de seu conhecimento.

Buscando a minimização das dificuldades enfrentadas pelos pesquisadores frente a atualização de seus conhecimentos, o trabalho também objetivou desenvolver uma abordagem que identificasse perfis similares de pesquisadores, com o intuito de gerar recomendações baseadas em seus perfis de forma personalizada. Para isso, trabalhou-se num métrica para calcular o perfil do pesquisador utilizando dados de seu currículo, através de um modelo de perfil específico que permitisse uma abrangência e adaptações frente a esse contexto. Desta forma, adaptou-se um modelo de perfil de pesquisadores para a obtenção da similaridade, não somente acrescentando elementos, mas também buscando o conteúdo textual desses elementos. Com esse processo, foi possível conseguir indicadores de similaridades quantitativas e qualitativas de seu perfil e, ainda, através do indicativo de grau de relevância dos termos do perfil, permitiu-se o acompanhamento das mudanças de preferências dos pesquisadores e seus termos relacionados ao seu currículo.

Como resultado do trabalho, desenvolveu-se uma solução computacional para realizar a coleta dos dados dos currículos, obtidos da Plataforma Lattes, implementando os cálculos necessários para os índices de similaridades propostos, o que permitiu aplicar as funções de recomendações propostas para os pesquisadores. Igualmente, uma ferramenta para a busca e visualização de tais recomendações foi implementada. Através de experimentos foi averiguado que a abordagem proposta tem boa cobertura de recomendações e que o algoritmo

tem potencial de descoberta de novos conteúdos úteis ao pesquisador, de forma que o sistema consegue compreender as preferências do pesquisador através do seu currículo e personalizar as recomendações.

A abordagem proposta e aplicada por meio de experimentos teve como premissa ser abrangente e adaptável. Abrangente, pois é possível utilizá-la em diferentes contextos, nos quais sejam necessários similaridade de perfil de pesquisadores. Nesse trabalho, ela foi avaliada para recomendações, porém, pode ser utilizada para a busca de perfis similares e para a descoberta de indivíduos e avaliação dos perfis dos mesmos. A abordagem também é considerada adaptável, pois dependendo dos tipos de áreas de atuação e de conhecimento dos pesquisadores pode-se configurar os pesos dos elementos conforme suas peculiaridades. Da mesma forma que também é fácil adaptar as funções de recomendação para atingir novas situações que possam surgir.

Com base na fundamentação teórica e no processo desenvolvido neste trabalho, foram desenvolvidos experimentos, usando um conjunto de dados reais dos pesquisadores de oito áreas do conhecimento contidas na Plataforma Lattes. Essa pesquisa também apresentou uma abordagem de recomendação que trabalhou com a hipótese de que perfis similares tendem a ter preferências similares, trabalhando com uma similaridade dentro de vários elementos do currículo do pesquisador para definir os perfis dos pesquisadores. Também se verificou que essa similaridade poderia levar a obtenção de melhorias significativas na função de recomendação e poderia aumentar significativamente a personalização da recomendação, utilizando cada elemento do perfil para isto. O trabalho desenvolvido incluiu propostas para se considerar os aspectos temporais e de ponderação nos termos textuais dos elementos de perfil de pesquisador, os quais são usados como base para a geração dos índices de similaridade e abastecendo as estratégias de recomendação.

Diante do exposto, conclui-se que os objetivos deste trabalho foram atendidos, considerando que a abordagem produziu recomendações com boa cobertura, podendo aumentar a descoberta de conteúdo útil ao pesquisador e sem a interferência do mesmo atualizar suas preferências, apenas utilizando dados de seu currículo.

Assim, o trabalho tem as seguintes contribuições evidenciadas: (i) adaptações no modelo de perfil Rep-Model, para o que mesmo envolva não somente a reputação do pesquisador como também capacidade de determinar a similaridade entre os perfis dos pesquisadores; (ii) uma métrica para calcular os índices de similaridade quantitativas e qualitativas, de forma local ou global, utilizando os pesos do modelo de perfil. Tal métrica é independente do modelo de perfil de pesquisador podendo ser utilizada em outros contextos;

(iii) um indicador do grau de relevância dos termos junto ao perfil, o qual consegue acompanhar as mudanças das preferências com o passar do tempo; (iv) a definição de regras de recomendações que utilizam os índices de similaridade, de forma que as mesmas possibilitem fáceis adaptações.

Cabe ressaltar que o tema abordado no trabalho é extenso e de grande interesse, permitindo que as pesquisas sobre o tema possam ter continuidade. Durante a realização do trabalho vislumbrou-se a possibilidade de vários trabalhos futuros, a saber: (i) realizar experimentos com pesos diferentes do modelo de perfil, levando-se em consideração as peculiaridades de cada área do conhecimento; (ii) realizar experimentos para todas as áreas de pesquisas do CNPq e CAPES; (iii) realizar experimentos com pesquisadores internacionais, utilizando dados da DBLP; (iv) estudar novos tipos de recomendações personalizadas, baseadas em outros elementos do perfil do pesquisador não desenvolvidos nesse trabalho; (v) estudar novas recomendações baseadas no grau de relevância dos termos; (vi) permitir ao usuário avaliar as recomendações de forma online, se manifestando sobre a relevância das recomendações que recebeu; (vii) acrescentar mais elementos no modelo de perfil para maiores comparações, aumentando assim a personalização dos conteúdo sugeridos aos usuário.

REFERÊNCIAS

[1] CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. De. Application of Scientific Metrics to Evaluate Academic Reputation in Different Research Areas. *Proceedings of XXXIV International Conference on Computational Science (ICCS 2013)*. 2013. p. 2778–2788.

[2] LIMA, H. *et al.* Assessing the profile of top Brazilian computer science researchers. *Scientometrics*. vol. 103, nº 3. p. 879–896. jun. 2015.

[3] SUGIYAMA, K.; KAN, M.-Y. Exploiting potential citation papers in scholarly paper recommendation. *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries - JCDL '13*. 2013. p. 153.

[4] HANNEL, K. *et al.* Qualificação de pesquisadores por área da Ciência da Computação com base em uma ontologia de perfil. *V Congr. Iberoam. Telemática. CITA 2009*. p. 88–95. 2009.

[5] CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. De. An Adaptive Approach for Identifying Reputation of Researchers. *IADIS Int. Conf. WWW/Internet, Marid*. 2012.

[6] LEE, J.; LEE, K.; KIM, J. Personalized Academic Research Paper Recommendation System. *arXiv Prepr. arXiv1304.5457*. p. 1–8. 2013.

[7] GOLLAPALLI, S. Das; MITRA, P.; GILES, C. L. Similar researcher search in academic environments. *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries - JCDL '12*. 2012. p. 167.

[8] MENA-CHALCO, J. P.; DIGIAMPIETRI, L. A; CESAR-JR, R. M. Caracterizando as redes de coautoria de currículos Lattes. *BraSNAM - Brazilian Work. Soc. Netw. Anal. Min.* p. 12. 2012.

[9] HONG, K.; JEON, H.; JEON, C. Personalized Research Paper Recommendation System using Keyword Extraction Based on UserProfile. *J. Converg. Inf. Technol.* vol. 8, nº 16. p. 106–116. 2013.

[10] TRAJKOVA, J.; GAUCH, S. Improving Ontology-based User Profiles. *Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*. 2004. p. 380–390.

[11] CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. De. Comparing the Reputation of Researchers Using a Profile Model and Scientific Metrics. *Proceedings of XIII IEEE International Conference on Computer and Information Technology (CIT 2013)*. 2013. p. 353–359.

[12] MIDDLETON, S. E. *et al.* Ontological User Profiling in Recommender Systems. *ACM Trans. Inf. Syst.* vol. 22, nº 1. p. 54–88. 2004.

[13] GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing? *Int. J. Hum. Comput. Stud.* vol. 43, nº 5. p. 907–928. 1995.

[14] BEEL, J. *et al.* Research-paper recommender systems : a literature survey. *Int. J. Digit. Libr.* nº February 2014. 2015.

[15] MONTANER, M. Collaborative recommender agents based on case-based reasoning and trust. PhD Thesis. Universitat de Girona 2003. 2003.

[16] SYED, H. H.; ANDRITSOS, P. User Preference Modeling - A Survey. *Tech. Rep. DIT-07-060*. 2007.

[17] WAINER, J. *et al.* Empirical evaluation in Computer Science research published by ACM. *Inf. Softw. Technol.* vol. 51, nº 6. p. 1081–1085. jun. 2009.

[18] WAINER, J.; XAVIER, E. C.; BEZERRA, F. Scientific production in Computer Science: A comparative study of Brazil and other countries. *Scientometrics*. vol. 81, nº

2. p. 535–547. nov. 2009.
[19] WAINER, J.; VIEIRA, P. Correlations between bibliometrics and peer evaluation for all disciplines: the evaluation of Brazilian scientists. *Scientometrics*. vol. 96, nº 2. p. 395–410. ago. 2013.
[20] HANNEL, K.; WARPECHOWSKI, M.; LIMA, J. V. De. Modelo para Identificar a Qualificação de Pesquisador nas Áreas da Ciência da Computação. *An. do XXVIII Congr. da SBC*. p. 147–156. 2008.
[21] VIVIAN, G. R.; CERVI, C. R. Utilizando Técnicas de Data Science para Definir o Perfil do Pesquisador Brasileiro da Área de Ciência da Computação. *Anais da XII ERBD*. 2016. p. 108–117.
[22] NASCIMENTO, C. *et al.* A source independent framework for research paper recommendation. *Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries - JCDL '11*. 2011. p. 297.
[23] MENA-CHALCO, J. P.; JUNIOR, R. M. C. scriptLattes: an open-source knowledge extraction system from the Lattes platform. *J. Brazilian Comput. Soc.* vol. 15, nº 4. p. 31–39. 2009.
[24] MAGALHÃES, J. L. *et al.* Extração e tratamento de dados na base lattes para identificação de core competencies em dengue. *Informação & Informação*. vol. 19, nº 3. p. 30. ago. 2014.
[25] GIORDANO, D. M.; BRUNING, E.; BORDIN, A. S. Uso do scriptLattes e Gephi na Análise da Colaboração Científica. *Comput. BEACH 2015*. p. 239–248. 2015.
[26] LEY, M. DBLP - Some Lessons Learned. *Site DBLP*. 2009.
[27] LEE, D. *et al.* Are Your Citations Clean? *Commun. ACM*. vol. 50, nº 12. p. 33–38. 2007.
[28] LEY, M.; REUTHER, P. Maintaining an Online Bibliographical Database: The Problem of Data Quality. *EGC2006 Actes des sixièmes journées Extr. Gest. des Connaissances*. vol. RNTI-E-6. p. 5–10. 2006.
[29] LAENDER, A. H. F. *et al.* Assessing the research and education quality of the top Brazilian Computer Science graduate programs. *ACM SIGCSE Bull.* vol. 40, nº 2. p. 135. jun. 2008.
[30] GUGEL, J. *et al.* Uma Ferramenta Para Análise Quantitativa da Produção Científica de Pesquisadores. *VII Esc. Reg. Banco Dados*. p. 1–10. 2011.
[31] WANGENHEIM, C. G. Von; WANGENHEIM, A. Von; RATEKE, T. *Raciocínio baseado em casos*. 2ª edição. Bookess Editora, 2013.
[32] ZHANG, M.; HURLEY, N. Novel item recommendation by user profile partitioning. *Proc. - 2009 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2009*. vol. 1. p. 508–515. 2009.
[33] RICCI, F.; ROKACH, L.; SHAPIRA, B. Introduction to Recommender Systems Handbook. in *Recommender Systems Handbook*. vol. 532. Boston, MA: Springer US. 2011. p. 1–35.
[34] SOUZA, R. G. D. De. Sistemas de Recomendação: Aplicando Sistemas de Recomendação em Situações Práticas. 28-maio-2014. Available at: <https://www.ibm.com/developerworks/br/local/data/sistemas_recomendacao/>. Acessado: 08 nov. 2017.
[35] NEWMAN, M. E. J. Coauthorship networks and patterns of scientific collaboration. *Proc. Natl. Acad. Sci.* vol. 101, nº Supplement 1. p. 5200–5205. abr. 2004.
[36] WASSERMAN, S.; FAUST, K. Social network analysis: methods and applications. *New York Cambridge Univ.* 1994.
[37] TANG, J. *et al.* ArnetMiner: Extraction and Mining of an Academic Social Network. *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discov. data Min.* p. 1193–1194. 2008.
[38] TANG, J. *et al.* Extraction and Mining of an Academic Social Network. *Proceedings of*

the 17th International Conference on World Wide Web. 2008. p. 1193–1194.

[39] TANG, J.; ZHANG, D.; YAO, L. Social Network Extraction of Academic Researchers. *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*. 2007. p. 292–301.

[40] BRANDÃO, M. A; MORO, M. M. Recomendação de Colaboração em Redes Sociais Acadêmicas baseada na Afiliação dos Pesquisadores. *Simpósio Bras. Bancos Dados - SBBD 2012*. p. 73–80. 2012.

[41] BRANDÃO, M. A.; MORO, M. M.; ALMEIDA, J. M. Análise de Fatores Impactantes na Recomendação de Colaborações Acadêmicas Utilizando Projeto Fatorial. *Simpósio Bras. Bancos Dados - SBBD 2013*. p. 1–6. 2013.

[42] MAIA, M. D. F. S.; CAREGNATO, S. E. Co-autoria como indicador de redes de colaboração científica. *Perspect. em Ciência da Informação*. vol. 13, nº 2. p. 18–31. ago. 2008.

[43] JR, P. S. P.; LAENDER, A. H. F.; MORO, M. M. Análise da Rede de Coautoria do Simpósio Brasileiro de Bancos de Dados. *Simpósio Bras. Banco Dados - SBBD 2010*. nº 573871. 2010.

[44] BARBOSA, E. M. *et al.* VRRC: Uma Ferramenta Web para Visualização e Recomendação em Redes de Coautoria. *VIII Sessão Demos, Simpósio Bras. Banco Dados*. 2011.

[45] DINIZ, M. A. *et al.* CNARe: Uma Ferramenta Online para Análise de Redes Sociais de Co-autoria e Recomendações. *SBBD – Demos Appl. Sess.* p. 143–148. 2015.

[46] JACCARD, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. del la Société Vaudoise des Sci. Nat.* vol. 37. p. 547–579. 1901.

[47] VIVIAN, G. R.; CERVI, C. R.; ROVADOSKY, D. N. Using selection attribute algorithms from data mining to complement the rep-index. *IADIS Int. J. WWW/Internet*. p. 219–226. 2016.

[48] LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions, and reversals. 1966.

[49] HERLOCKER, J. L. *et al.* Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.* vol. 22, nº 1. p. 5–53. 2004.

[50] GE, M.; DELGADO-BATTENFELD, C.; JANNACH, D. Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. *Proc. fourth ACM Conf. Recomm. Syst. - RecSys '10*. p. 257. 2010.

[51] HUANG, W. *et al.* Recommending Citations: Translating Papers into References. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. 2012. p. 1910–1914.

APÊNDICE A - EXTRAÇÃO DOS DADOS DOS PESQUISADORES

Aqui demonstra-se os passos realizador para a extração dos dados abertos da Plataforma Lattes, através de imagens feitas do processo.

Figura 17: Site Plataforma Lattes extração dos dados de janeiro de 2017.

Figura 18: Download dos dados (números identificadores) e da padronização dos mesmos.

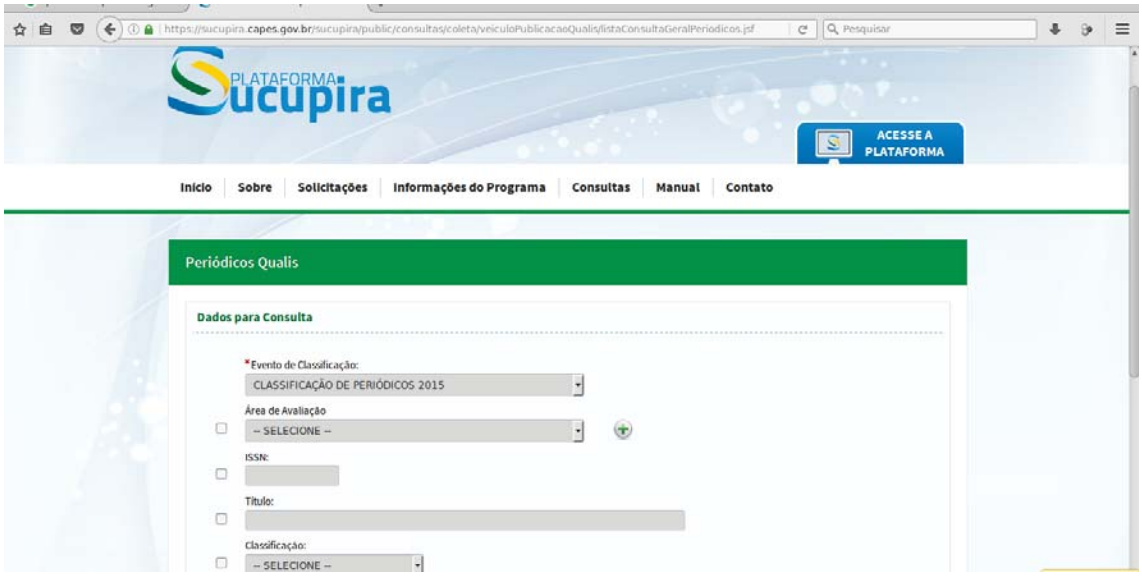


Figura 19: Consulta a Plataforma Sucupira extração dos dados de classificação dos periódicos Qualis. Pesquisa feita em janeiro de 2017, dados referentes ao ano de 2015.

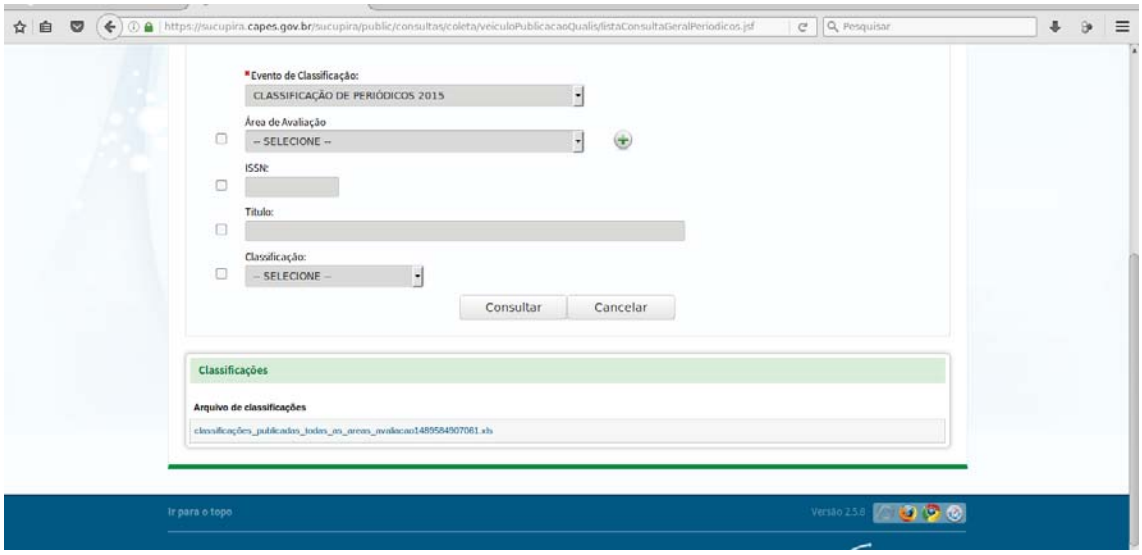


Figura 20: Arquivo com a classificação de todas as áreas.

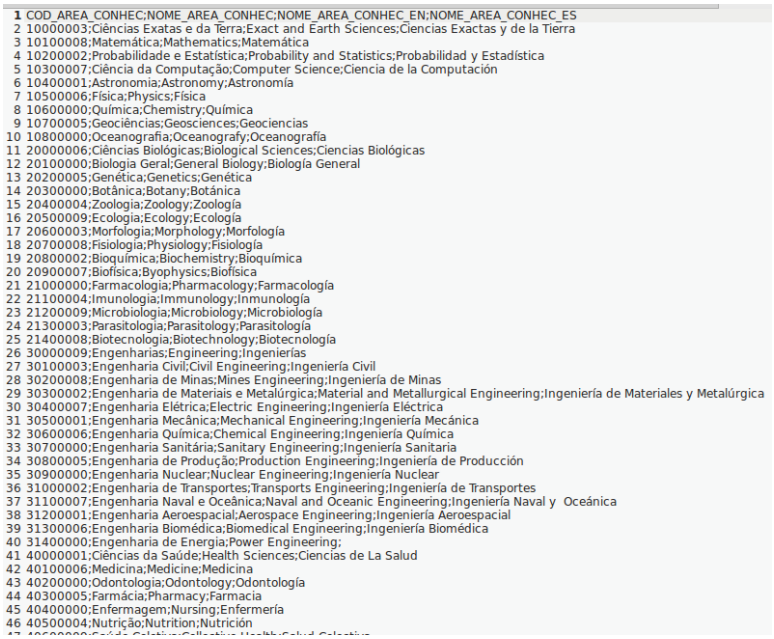
```
<?xml version="1.0" encoding="ISO-8859-1"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
           elementFormDefault="qualified"
           attributeFormDefault="unqualified">
 <xs:element name="CURRICULO-VITAE">
  <xs:complexType>
   <xs:sequence>
    <xs:element ref="DADOS-GERAIS"/>
    <xs:element minOccurs="0" ref="PRODUCAO-BIBLIOGRAFICA"/>
    <xs:element minOccurs="0" ref="PRODUCAO-TECNICA"/>
    <xs:element minOccurs="0" ref="OUTRA-PRODUCAO"/>
    <xs:element minOccurs="0" ref="DADOS-COMPLEMENTARES"/>
   </xs:sequence>
   <xs:attribute name="SISTEMA-ORIGEM-XML" use="required"/>
   <xs:attribute name="NUMERO-IDENTIFICADOR"/>
   <xs:attribute name="FORMATO-DATA-ATUALIZACAO" default="DDMMAAAA">
    <xs:simpleType>
     <xs:restriction base="xs:NMTOKEN">
      <xs:enumeration value="DDMMAAAA"/>
     </xs:restriction>
    </xs:simpleType>
   </xs:attribute>
   <xs:attribute name="DATA-ATUALIZACAO"/>
   <xs:attribute name="FORMATO-HORA-ATUALIZACAO" default="HHMMSS">
    <xs:simpleType>
     <xs:restriction base="xs:NMTOKEN">
      <xs:enumeration value="HHMMSS"/>
     </xs:restriction>
    </xs:simpleType>
   </xs:attribute>
   <xs:attribute name="HORA-ATUALIZACAO"/>
  </xs:complexType>
 </xs:element>
 <xs:element name="DADOS-GERAIS">
  <xs:complexType>
   <xs:sequence>
    <xs:element minOccurs="0" ref="RESUMO-CV"/>
    <xs:element minOccurs="0" ref="OUTRAS-INFORMACOES-RELEVANTES"/>
    <xs:element minOccurs="0" ref="ENDERECO"/>
    <xs:element minOccurs="0" ref="FORMACAO-ACADEMICA-TITULACAO"/>
    <xs:element minOccurs="0" ref="ATUACOES-PROFISSIONAIS"/>
    <xs:element minOccurs="0" ref="AREAS-DE-ATUACAO"/>
    <xs:element minOccurs="0" ref="IDIOMAS"/>
    <xs:element minOccurs="0" ref="PREMIOS-TITULOS"/>
   </xs:sequence>
   <xs:attribute name="NOME-COMPLETO" use="required"/>
```

Figura 21: Parte do arquivo XSD (CurriculoLattes.xsd)[22].

```
NUMERO_IDENTIFICADOR;PAIS_NACIONALIDADE_ISO3;DATA_ATUALIZACAO;COD_AREA_CONHEC;COD_NIVEL_FORMACAO
77397926978835 57;BRA;29/09/2009 18:36:17;;2
88719545171955 36;BRA;14/10/2015 20:17:54;60400005;C
96053208352062 869;BRA;28/03/2013 17:58:06;40400000;1
36327567027704 09;BRA;24/04/2016 18:43:16;70800006;C
22198162405031 60;BRA;08/10/2015 14:44:07;;C
48391363007749 44;BRA;18/10/2013 13:44:36;;C
36903081963787 42;BRA;21/09/2015 17:07:51;;2
30451772461324 651;BRA;01/07/2016 01:14:46;30400007;4
81514193911545 96;BRA;19/02/2015 19:35:19;30100003;C
74266724633135 34;BRA;18/04/2016 11:18:59;30800005;C
98558072828658 47;BRA;24/03/2008 16:07:26;80300006;
08421054109990 86;BRA;02/02/2014 13:55:10;60900008;1
01148634872496 29;BRA;31/10/2016 17:36:18;10600000;C
15484245500323 35;BRA;08/09/2014 12:27:58;;1
93730492961629 06;BRA;29/03/2016 13:37:08;;2
15256721837279 61;BRA;02/11/2016 13:37:28;10300007;2
17990682854179 86;BRA;02/01/2017 14:35:12;10700005;3
83743035552572 56;BRA;15/07/2016 14:02:36;50500007;2
20230003466612 50;BRA;03/12/2013 13:12:52;70700001;C
86236866201483 63;BRA;03/04/2015 19:37:08;;3
26311040982872 01;BRA;28/10/2015 17:49:16;;C
63146355131645 30;BRA;10/12/2016 06:41:21;70500002;1
99098925112506 86;BRA;16/04/2012 22:32:01;40100006;4
43114381762375 09;BRA;06/09/2011 13:54:17;;C
99318292420670 15;BRA;30/05/2016 21:41:47;10300007;C
96468001765017 40;BRA;03/10/2016 08:25:48;10500006;2
78985195332771 32;BRA;26/02/2015 11:05:32;10300007;2
23405413702366 79;BRA;05/01/2017 01:57:06;61000000;1
00883291928801 34;BRA;28/01/2013 16:29:24;;A
46016538902166 03;BRA;12/07/2012 09:58:55;70800006;1
98652089526493 28;BRA;14/07/2015 20:57:18;;2
20667852875821 65;BRA;01/04/2006 00:00:00;;
21819468956430 77;BRA;31/03/2016 10:10:21;;C
98245612704188 34;BRA;21/09/2006 00:00:00;;
44620249368347 00;BRA;14/03/2016 21:48:24;40900002;2
25779400482087 89;BRA;20/01/2016 19:41:07;40800008;2
24497003725731 31;BRA;27/12/2016 22:47:22;80100007;4
02668436144046 42;BRA;11/09/2009 10:41:30;;B
33917793543555 85;CZE;21/06/2009 20:13:19;40100006;3
51918580207114 08;BRA;08/12/2009 14:10:45;;
65962175544017 86;BRA;19/09/2016 16:14:10;70600007;2
58726936172449 18;BRA;18/11/2005 00:00:00;60200006;1
00425676340401 93;BRA;03/03/2015 10:59:23;;1
07833976474950 37;BRA;13/05/2014 21:21:59;30600006;C
32891603302972 69;BRA;06/12/2016 14:31:19;20200005;C
```

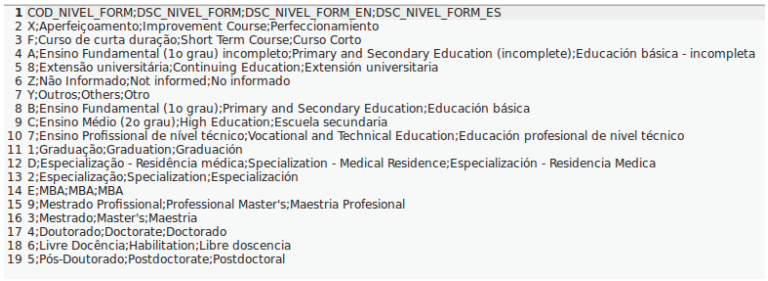
Figura 22: Parte do arquivo numero_identificador_lattes_20170108.csv.

[22] O arquivo CurriculoLattes.xsd contém as regras de validação dos documentos curriculo.xml os quais contém os dados dos currículo do pesquisador. A linguagem XML Schema é baseada no formato XML e é uma alternativa ao DTD.



```
COD_AREA_CONHEC;NOME_AREA_CONHEC;NOME_AREA_CONHEC_EN;NOME_AREA_CONHEC_ES
10000003;Ciências Exatas e da Terra;Exact and Earth Sciences;Ciencias Exactas y de la Tierra
10100008;Matemática;Mathematics;Matemática
10200002;Probabilidade e Estatística;Probability and Statistics;Probabilidad y Estadística
10300007;Ciência da Computação;Computer Science;Ciencia de la Computación
10400001;Astronomia;Astronomy;Astronomía
10500006;Física;Physics;Física
10600000;Química;Chemistry;Química
10700005;Geociências;Geosciences;Geociencias
10800000;Oceanografia;Oceanografy;Oceanografía
20000006;Ciências Biológicas;Biological Sciences;Ciencias Biológicas
20100000;Biologia Geral;General Biology;Biología General
20200005;Genética;Genetics;Genética
20300000;Botânica;Botany;Botánica
20400004;Zoologia;Zoology;Zoología
20500009;Ecologia;Ecology;Ecología
20600003;Morfologia;Morphology;Morfología
20700008;Fisiologia;Physiology;Fisiología
20800002;Bioquímica;Biochemistry;Bioquímica
20900007;Biofísica;Byophysics;Biofísica
21000000;Farmacologia;Pharmacology;Farmacología
21100004;Imunologia;Immunology;Inmunología
21200009;Microbiologia;Microbiology;Microbiología
21300003;Parasitologia;Parasitology;Parasitología
21400008;Biotecnologia;Biotechnology;Biotecnología
30000009;Engenharias;Engineering;Ingenierías
30100003;Engenharia Civil;Civil Engineering;Ingeniería Civil
30200008;Engenharia de Minas;Mines Engineering;Ingeniería de Minas
30300002;Engenharia de Materiais e Metalúrgica;Material and Metallurgical Engineering;Ingeniería de Materiales y Metalúrgica
30400007;Engenharia Elétrica;Electric Engineering;Ingeniería Eléctrica
30500001;Engenharia Mecânica;Mechanical Engineering;Ingeniería Mecánica
30600006;Engenharia Química;Chemical Engineering;Ingeniería Química
30700000;Engenharia Sanitária;Sanitary Engineering;Ingeniería Sanitaria
30800005;Engenharia de Produção;Production Engineering;Ingeniería de Producción
30900000;Engenharia Nuclear;Nuclear Engineering;Ingeniería Nuclear
31000002;Engenharia de Transportes;Transports Engineering;Ingeniería de Transportes
31100007;Engenharia Naval e Oceânica;Naval and Oceanic Engineering;Ingeniería Naval y  Oceánica
31200001;Engenharia Aeroespacial;Aerospace Engineering;Ingeniería Aeroespacial
31300006;Engenharia Biomédica;Biomedical Engineering;Ingeniería Biomédica
31400000;Engenharia de Energia;Power Engineering;
40000001;Ciências da Saúde;Health Sciences;Ciencias de La Salud
40100006;Medicina;Medicine;Medicina
40200000;Odontologia;Odontology;Odontología
40300005;Farmácia;Pharmacy;Farmacia
40400000;Enfermagem;Nursing;Enfermería
40500004;Nutrição;Nutrition;Nutrición
```

Figura 23: Parte do arquivo tab_area_conhecimento_20170108.csv.



```
COD_NIVEL_FORM;DSC_NIVEL_FORM;DSC_NIVEL_FORM_EN;DSC_NIVEL_FORM_ES
X;Aperfeiçoamento;Improvement Course;Perfeccionamiento
F;Curso de curta duração;Short Term Course;Curso Corto
A;Ensino Fundamental (1o grau) incompleto;Primary and Secondary Education (incomplete);Educación básica - incompleta
8;Extensão universitária;Continuing Education;Extensión universitaria
Z;Não Informado;Not informed;No informado
Y;Outros;Others;Otro
B;Ensino Fundamental (1o grau);Primary and Secondary Education;Educación básica
C;Ensino Médio (2o grau);High Education;Escuela secundaria
7;Ensino Profissional de nível técnico;Vocational and Technical Education;Educación profesional de nivel técnico
1;Graduação;Graduation;Graduación
D;Especialização - Residência médica;Specialization - Medical Residence;Especialización - Residencia Medica
2;Especialização;Specialization;Especialización
E;MBA;MBA;MBA
9;Mestrado Profissional;Professional Master's;Maestria Profesional
3;Mestrado;Master's;Maestria
4;Doutorado;Doctorate;Doctorado
6;Livre Docência;Habilitation;Libre doscencia
5;Pós-Doutorado;Postdoctorate;Postdoctoral
```

Figura 24: Parte do arquivo tab_nivel_formacao_20170108.csv.

```
ISSN;Título;Área de Avaliação;Estrato
2318-4965;ABCS HEALTH SCIENCES;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B3
1980-4814;ABCUSTOS (SÃO LEOPOLDO, RS);ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B4
1012-8255;ACADEMIA (CARACAS);ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B1
1042-9670;ACADEMIC PSYCHIATRY;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;A1
0001-4273;ACADEMY OF MANAGEMENT JOURNAL;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;A1
1537-260X;ACADEMY OF MANAGEMENT LEARNING & EDUCATION;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;A1
2358-6559;ACANTO EM REVISTA;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B5
0001-4575;ACCIDENT ANALYSIS AND PREVENTION;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;A1
1368-0668;ACCOUNTING AUDITING & ACCOUNTABILITY JOURNAL;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
1744-9499;ACCOUNTING IN EUROPE (ONLINE);ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;A2
0951-3574;ACCOUNTING, AUDITING & ACCOUNTABILITY JOURNAL;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;A1
2152-2820;ACCOUNTING, ECONOMICS AND LAW: A CONVIVIUM;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B4
2317-0190;ACTA FISIÁTRICA;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B4
1980-5772;ACTA GEOGRÁFICA (UFRR);ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B3
2179-975X;ACTA LIMNOLOGICA BRASILIENSIA (ONLINE);ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
0102-6712;ACTA LIMNOLÓGICA BRASILIENSIA;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
1982-0194;ACTA PAUL DE ENFERM;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B1
1679-9291;ACTA SCIENTIARUM. HEALTH SCIENCES (IMPRESSO);ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
1807-8656;ACTA SCIENTIARUM. HUMAN AND SOCIAL SCIENCES;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
1679-7361;ACTA SCIENTIARUM. HUMAN AND SOCIAL SCIENCES (IMPRESSO);ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISI
1806-2563;ACTA SCIENTIARUM. TECHNOLOGY (IMPRESSO);ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
1807-8664;ACTA SCIENTIARUM. TECHNOLOGY (ONLINE);ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
1850-2032;ACTAS DE DISEÑO;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B4
2215-3535;ACTUALIDADES EN PSICOLOGÍA;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
1969-6728;ACTUEL MARX;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B1
1896-9461;AD AMERICAM;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B5
1518-9929;ADM. MADE (UNIVERSIDADE ESTÁCIO DE SÁ);ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
2237-5139;ADM.MADE;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
0095-3997;ADMINISTRATION & SOCIETY;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;A1
2316-7548;ADMINISTRAÇÃO DE EMPRESAS EM REVISTA;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B5
2175-5787;ADMINISTRAÇÃO PÚBLICA E GESTÃO SOCIAL;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
2358-0917;ADMINISTRAÇÃO: ENSINO E PESQUISA (RAEP);ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
2177-6083;ADMINISTRAÇÃO: ENSINO R PESQUISA;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
1983-7089;ADMpg: GESTÃO ESTRATÉGICA.;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B4
1679-9941;ADOLESCÊNCIA & SAÚDE (UERJ);ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B2
2042-4868;ADVANCE JOURNAL OF FOOD SCIENCE AND TECHNOLOGY;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;A2
2042-4876;ADVANCED JOURNAL OF FOOD SCIENCE AND TECHNOLOGY;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;A2
1662-8985;ADVANCED MATERIALS RESEARCH (ONLINE);ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B1
0882-6110;ADVANCES IN ACCOUNTING;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B1
0098-9258;ADVANCES IN CONSUMER RESEARCH;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;A2
2164-2648;ADVANCES IN INFECTIOUS DISEASES;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B4
2194-5357;ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B1
1474-7979;ADVANCES IN INTERNATIONAL MARKETING;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B1
1983-8611;ADVANCES IN SCIENTIFIC AND APPLIED ACCOUNTING;ADMINISTRAÇÃO PÚBLICA E DE EMPRESAS, CIÊNCIAS CONTÁBEIS E TURISMO;B1
```

Figura 25: Parte do arquivo classificacoes_publicadas_todas_as_areas_avaliacao.csv[23].

Depois de selecionado os dados para download foram correlacionados, através de uma consulta no site da plataforma, os identificadores dos currículos com os identificadores que são utilizados para consulta do currículo na web.

```php
    public function pegarLattesMenorIds()
    {
        $cont_line = 0;
        $arch = getcwd().'/data/lattes/csv/dados_identificador10300007.csv';
        $new_arch = getcwd().'/data/lattes/csv/dados_identificador_maior_menor10300007.csv';
        $f = fopen($new_arch,"w");
        $pointer = fopen($arch, "r");
        while (($dataset = fgetcsv($pointer, 4096, ";")) !== FALSE)
        {
            $data = array(
                "lni_numero_identificador" => $dataset[0]
                );
            if ($data["lni_numero_identificador"] != 'NUMERO_IDENTIFICADOR')
            {
                print $url_download = 'http://lattes.cnpq.br/'.$data["lni_numero_identificador"];
                print " - ";
                $html = new DOMDocument();
                @$html->loadHTMLFile($url_download);
                $xpath = new DOMXPath($html);
                $links = $xpath->query('//form//input');
                foreach ($links as $link)
                {
                    if ($link->getAttribute('id') == 'id')
                    {
                        print $data = $data["lni_numero_identificador"].';'.$link->getAttribute('value')."\n";
                        fwrite($f,$data);
                    }
                }
            }
            $cont_line++;
            print $cont_line."\n";
        }
        fclose ($pointer);
        fclose ($f);
        print "Quantidade de linhas lidas = $cont_line."."\n";
    }
```

Figura 26: Código em PHP para a busca e correlação dos identificadores.

[23] O arquivo contém os dados da classificação dos periódicos segundo a CAPES. Nele está contido em cada linha, o ISSN, o título do periódico, a área de avaliação e seu estrato (nível de qualidade).

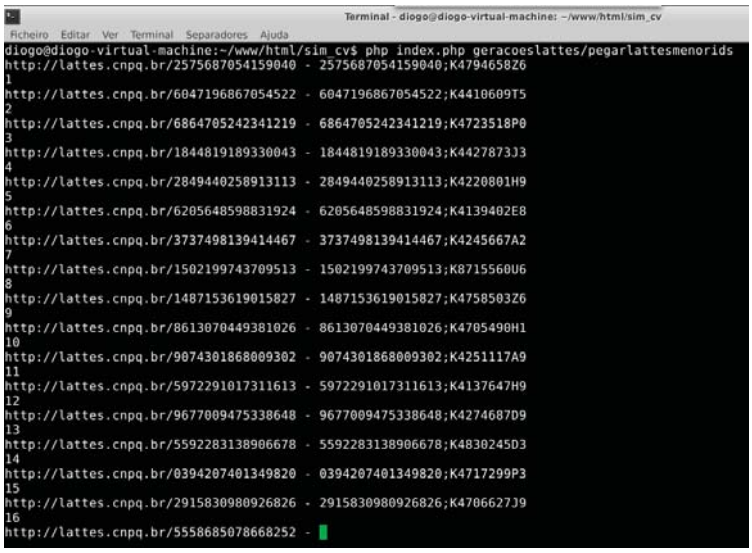


Figura 27: Execução do código para correlação dos identificadores.

```php
    public function downloadLattesXml()
    {
        $cont_line = 0;
        $url_download = 'buscacv.cnpq.br/buscacv/rest/download/curriculo/';
        $dir_arch = getcwd().'/data/geracoes/lattes/xml_zip10300007/';
        $arch = getcwd().'/data/lattes/csv/dados_identificador_maior_menor10300007.csv';
        $pointer = fopen($arch, "r");
        while (($dataset = fgetcsv($pointer, 4096, ";")) !== FALSE)
        {
            if ($dataset[1] != '')
            {
                $origem = $url_download.$dataset[1];
                $destino = $dir_arch;
                $result = shell_exec("wget -c -P $destino $origem");
            }
            $cont_line++;
            print $cont_line."\n";
        }
            fclose ($pointer);
            print "Quantidade de linhas lidas = $cont_line."."\n";
    }
```

Figura 28: Código PHP para leitura dos dados correlacionados e download do arquivo xml do currrículo.

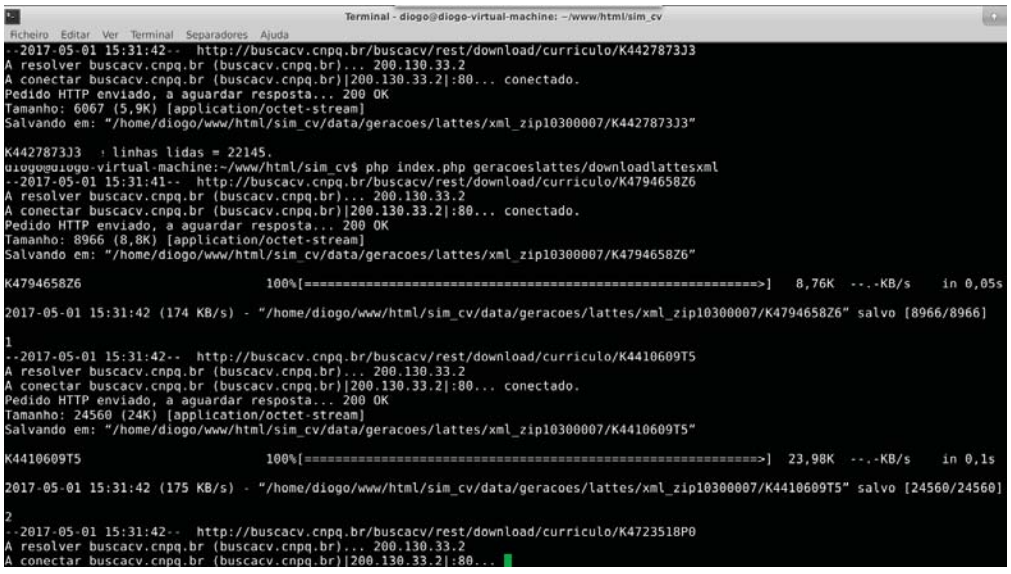


Figura 29: Execução do código de download.

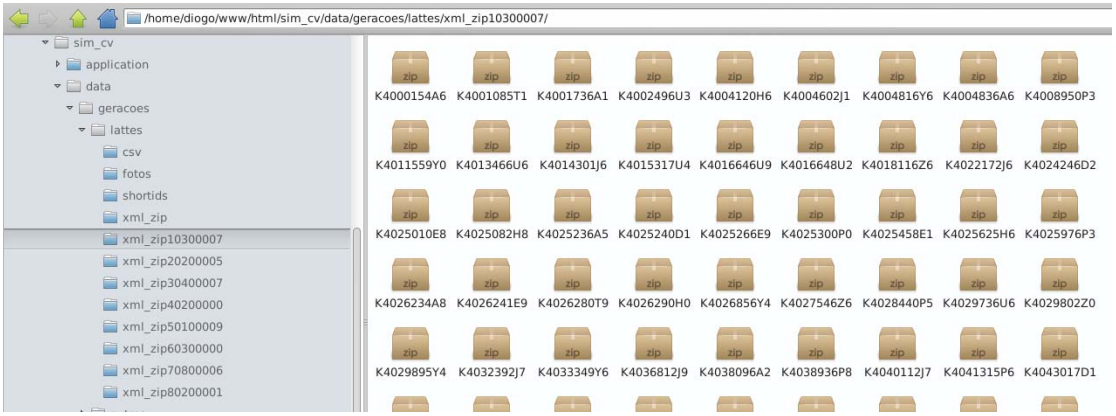


Figura 30: Parte dos arquivos baixados dos dados dos pesquisadores.

```
#!/bin/bash
for i in K*
do
        unzip  $i
        mv curriculo.xml /home/diogo/www/html/sim_cv/data/lattes/xml_10300007/$i.xml
done
```

Figura 31: Script para descompactar dados dos pesquisadores.

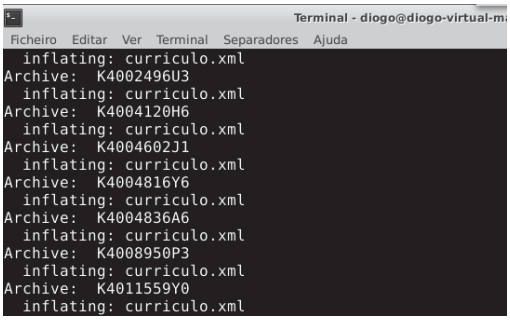


Figura 32: Execução do script para descompactar arquivos com os XML dos psequisadores.

Figura 33: Arquivos XMLs dos pesquisadores.

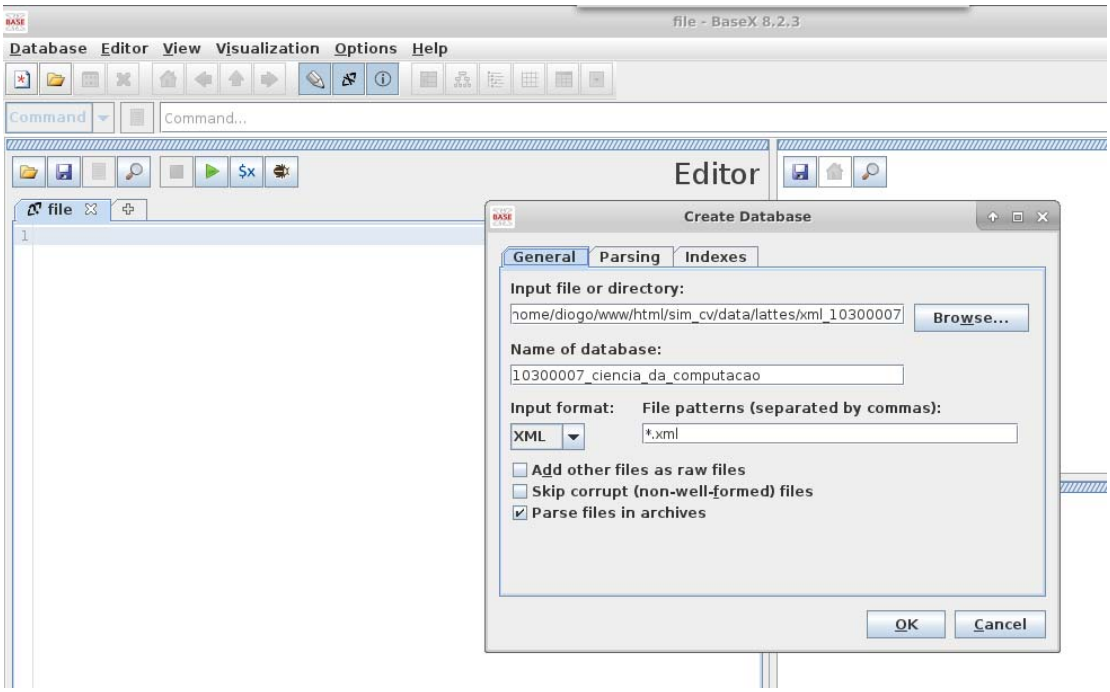


Figura 34: Criando uma base de dados XML com os arquivos descompactados do Lattes.

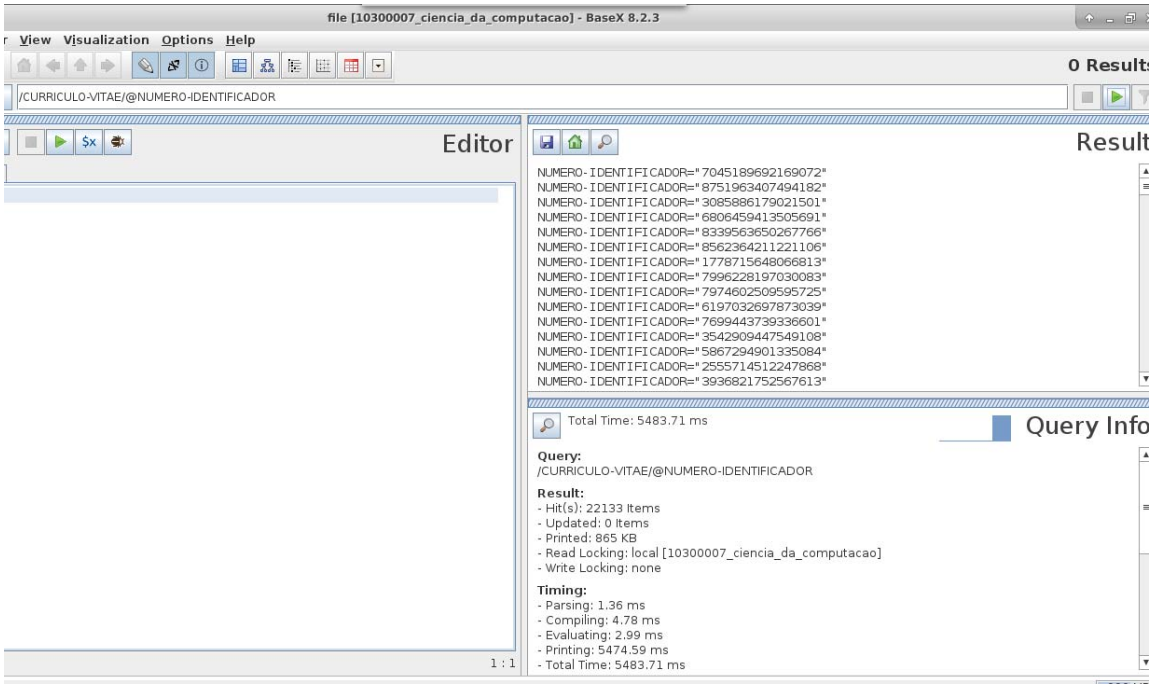


Figura 35: Busca pelos número identificadores dos arquivos XMLs.

APÊNDICE B - CARGA INICIAL DOS DADOS DOS PESQUISADORES

Aqui, demonstra-se os passos realizador de toda a carga necessária para montar o perfis dos pesquisadores, através de imagens feitas do processo, utilizando-se de uma ferramenta ETL, para popular as tabelas da base de dados relacional.

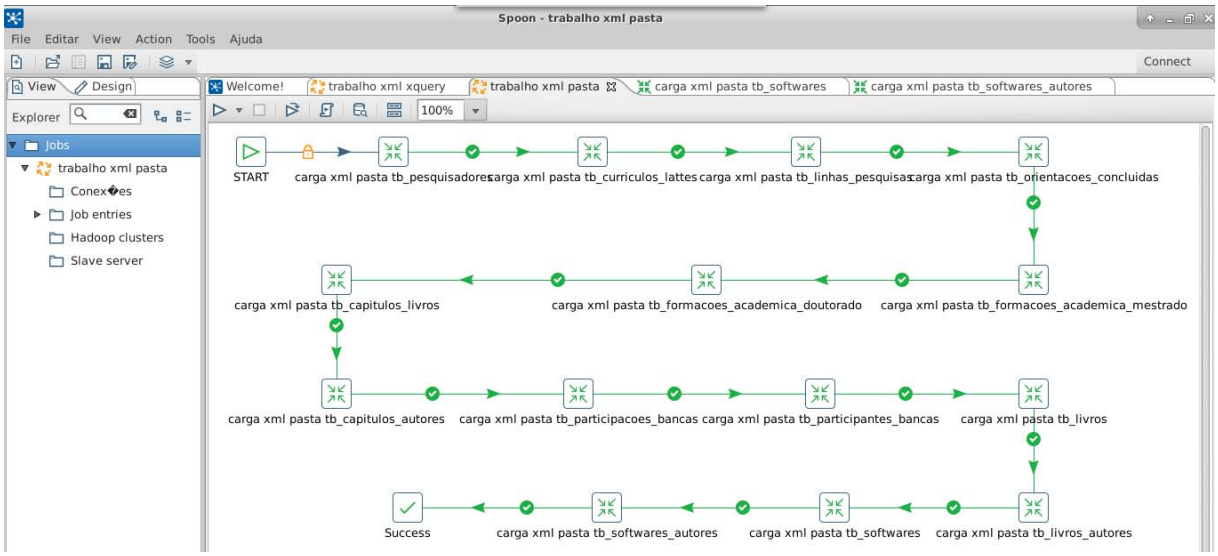


Figura 36: Visão geral do trabalho a ser executado pela ferramenta de ETL da carga dos dados.

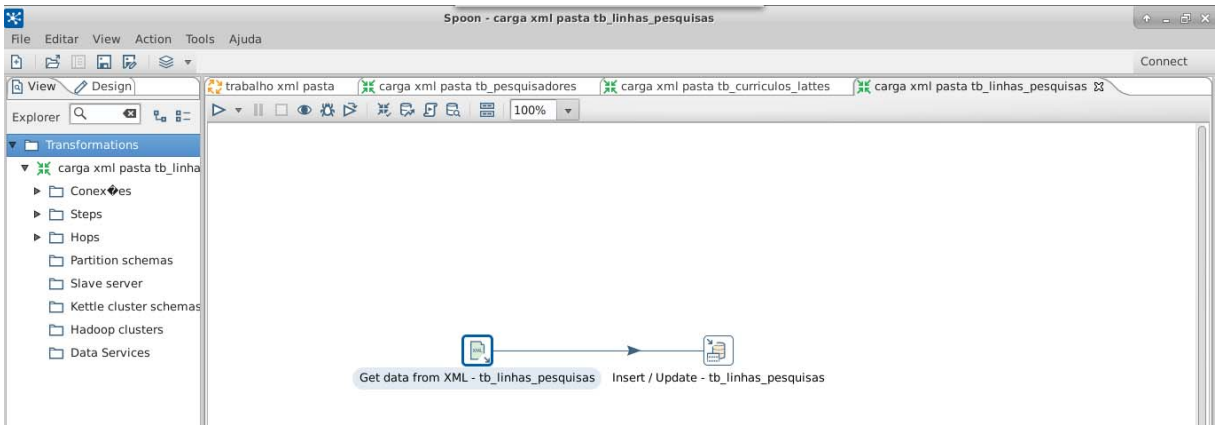


Figura 37: Visualização de um passo do trabalho de transformação dos dados, importando os arquivos XMLs, para dentro da tabela do banco de dados, realizado para cada tabela da base.

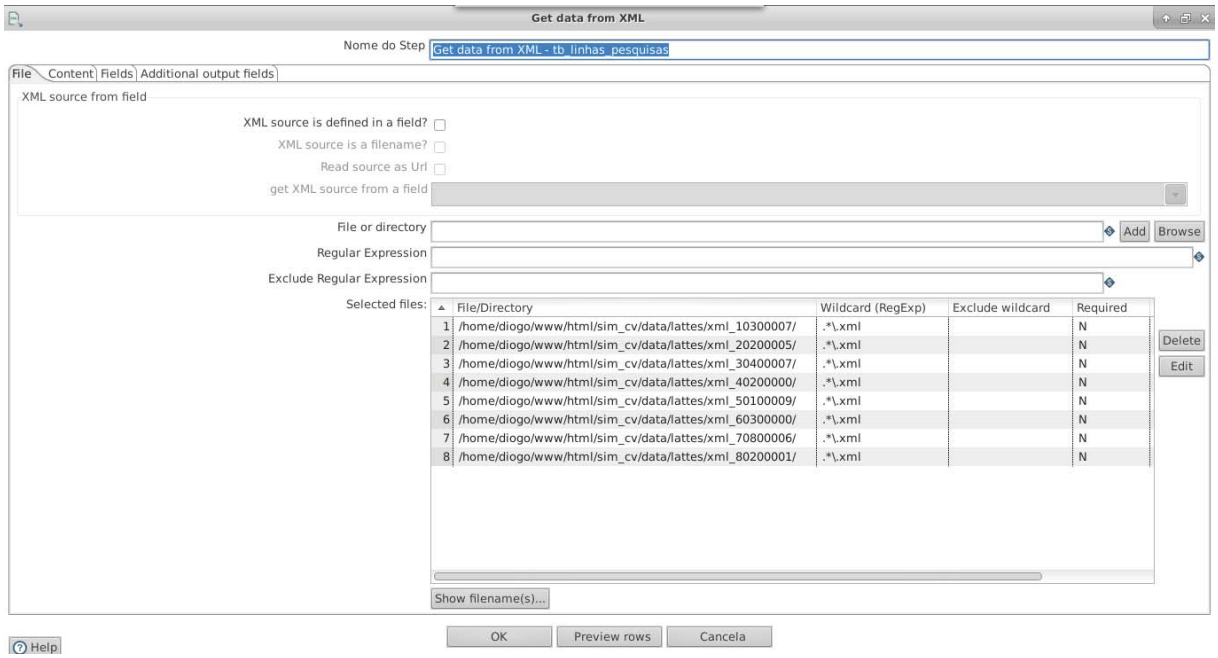


Figura 38: Configurando os diretórios dos dados dos pesquisadores baixados da Plataforma Lattes na ferramenta ETL.

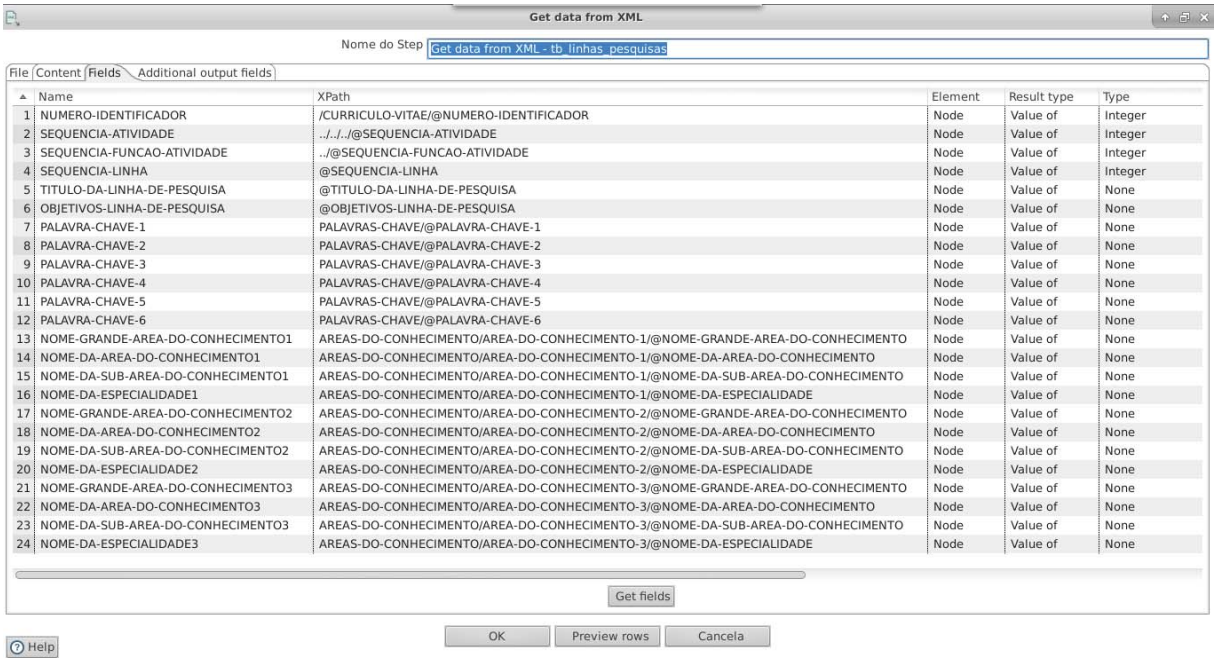


Figura 39: Configuração dos campos utilizados para importação dentro da ferramenta ETL.

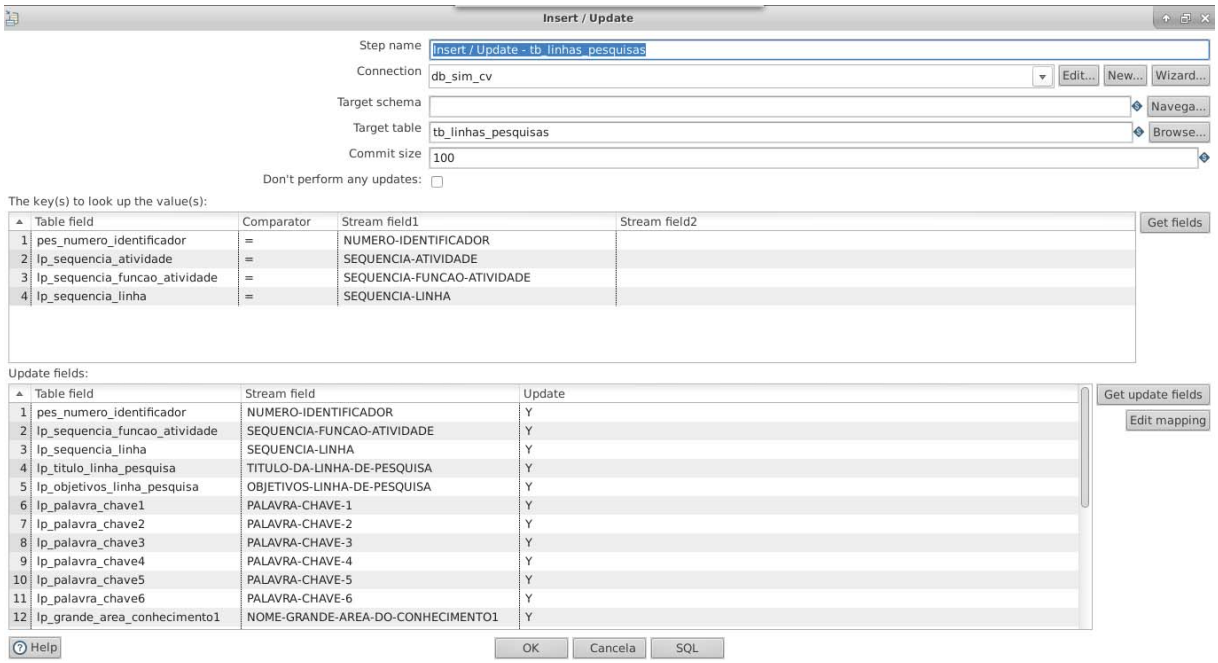


Figura 40: Correlação dos campos vindos do XML para os campos da tabela da base de dados, dentro da ferramenta ETL.

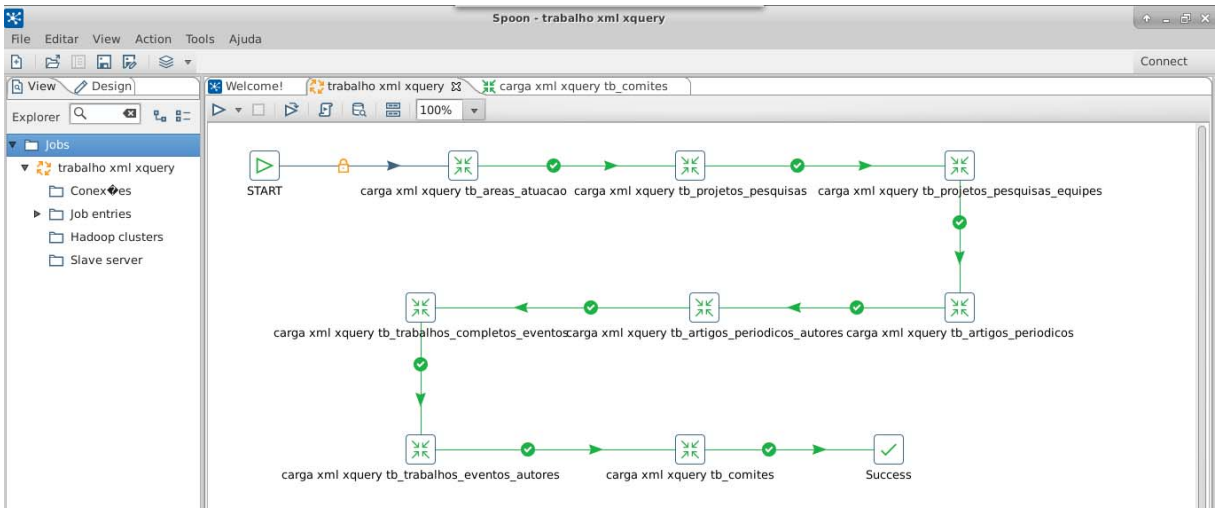


Figura 41: Visão geral do trabalho de carga dos dados exportados via software BaseX através de xQuery.

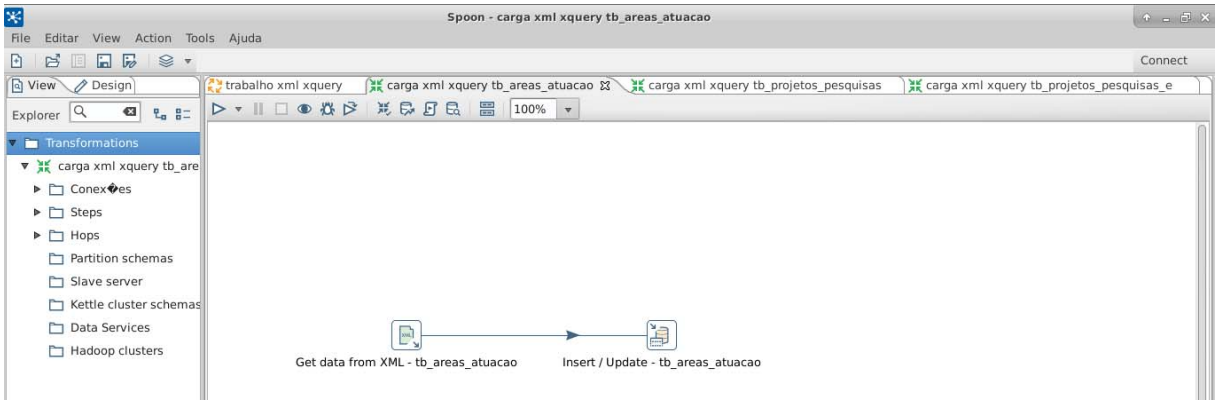


Figura 42: Transformação dos dados dos pesquisadores para dentro do banco de dados.

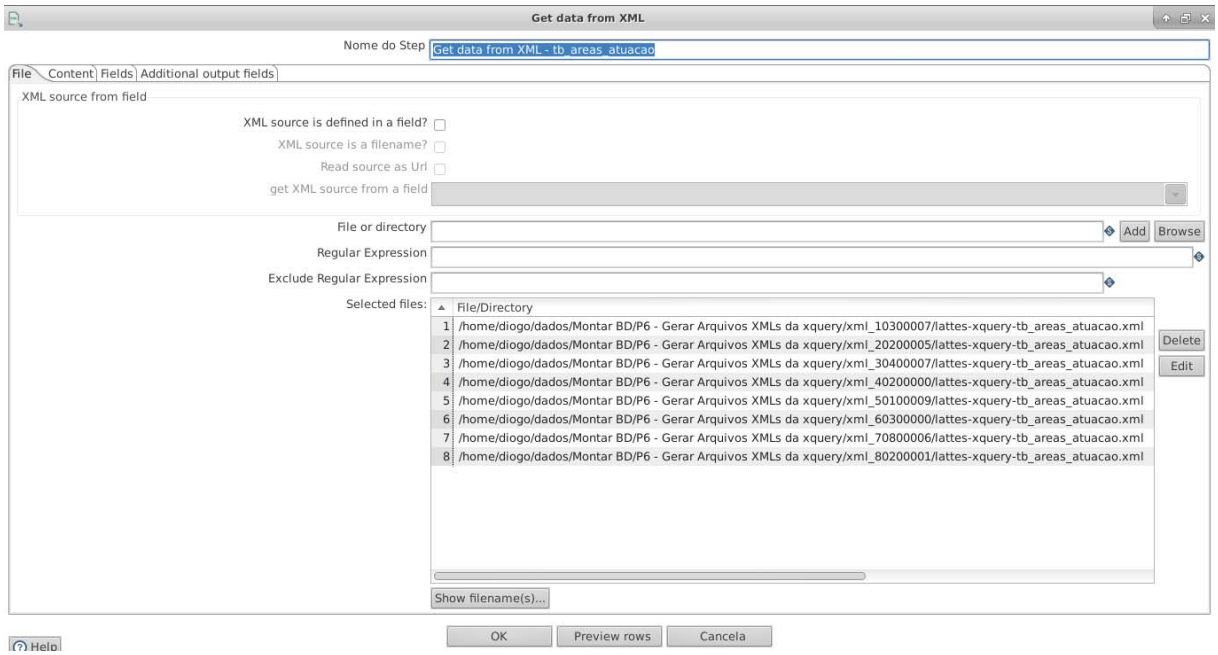


Figura 43: Diretórios dos dados exportados via xQuery sendo configurado na ferramenta ETL.

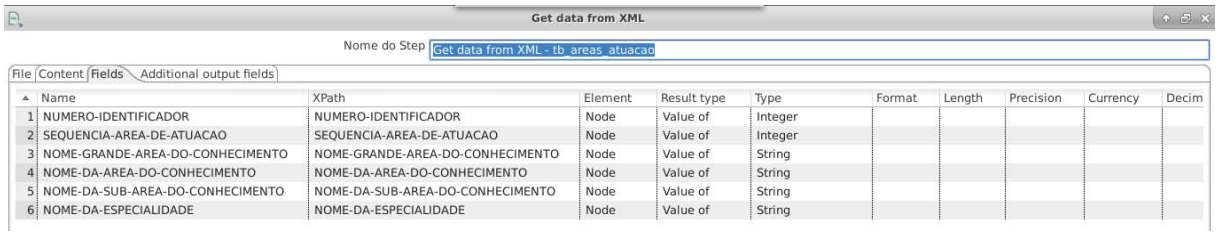
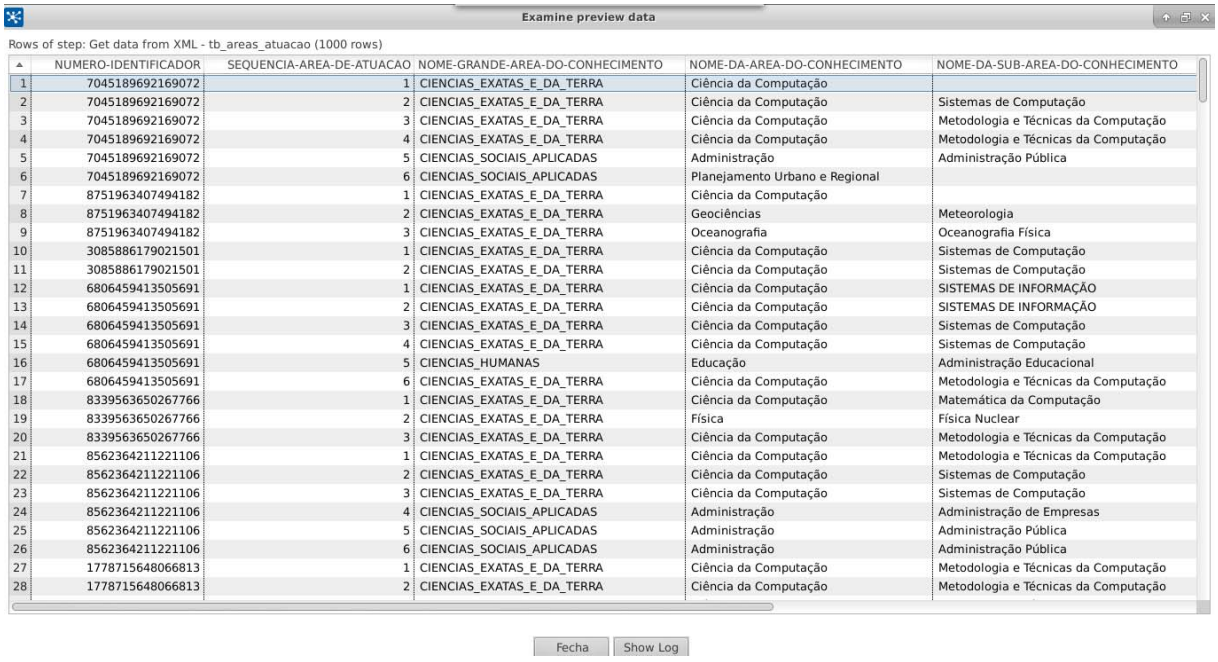


Figura 44: Configuração dos campos para importação.



	NUMERO-IDENTIFICADOR	SEQUENCIA-AREA-DE-ATUACAO	NOME-GRANDE-AREA-DO-CONHECIMENTO	NOME-DA-AREA-DO-CONHECIMENTO	NOME-DA-SUB-AREA-DO-CONHECIMENTO
1	7045189692169072	1	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	
2	7045189692169072	2	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Sistemas de Computação
3	7045189692169072	3	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Metodologia e Técnicas da Computação
4	7045189692169072	4	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Metodologia e Técnicas da Computação
5	7045189692169072	5	CIENCIAS_SOCIAIS_APLICADAS	Administração	Administração Pública
6	7045189692169072	6	CIENCIAS_SOCIAIS_APLICADAS	Planejamento Urbano e Regional	
7	8751963407494182	1	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	
8	8751963407494182	2	CIENCIAS_EXATAS_E_DA_TERRA	Geociências	Meteorologia
9	8751963407494182	3	CIENCIAS_EXATAS_E_DA_TERRA	Oceanografia	Oceanografia Física
10	3085886179021501	1	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Sistemas de Computação
11	3085886179021501	2	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Sistemas de Computação
12	6806459413505691	1	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	SISTEMAS DE INFORMAÇÃO
13	6806459413505691	2	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	SISTEMAS DE INFORMAÇÃO
14	6806459413505691	3	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Sistemas de Computação
15	6806459413505691	4	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Sistemas de Computação
16	6806459413505691	5	CIENCIAS_HUMANAS	Educação	Administração Educacional
17	6806459413505691	6	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Metodologia e Técnicas da Computação
18	8339563650267766	1	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Matemática da Computação
19	8339563650267766	2	CIENCIAS_EXATAS_E_DA_TERRA	Física	Física Nuclear
20	8339563650267766	3	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Metodologia e Técnicas da Computação
21	8562364211221106	1	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Metodologia e Técnicas da Computação
22	8562364211221106	2	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Sistemas de Computação
23	8562364211221106	3	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Sistemas de Computação
24	8562364211221106	4	CIENCIAS_SOCIAIS_APLICADAS	Administração	Administração de Empresas
25	8562364211221106	5	CIENCIAS_SOCIAIS_APLICADAS	Administração	Administração Pública
26	8562364211221106	6	CIENCIAS_SOCIAIS_APLICADAS	Administração	Administração Pública
27	1778715648066813	1	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Metodologia e Técnicas da Computação
28	1778715648066813	2	CIENCIAS_EXATAS_E_DA_TERRA	Ciência da Computação	Metodologia e Técnicas da Computação

Figura 45: Pré-visualização dos dados para carga na tabela do banco de dados.

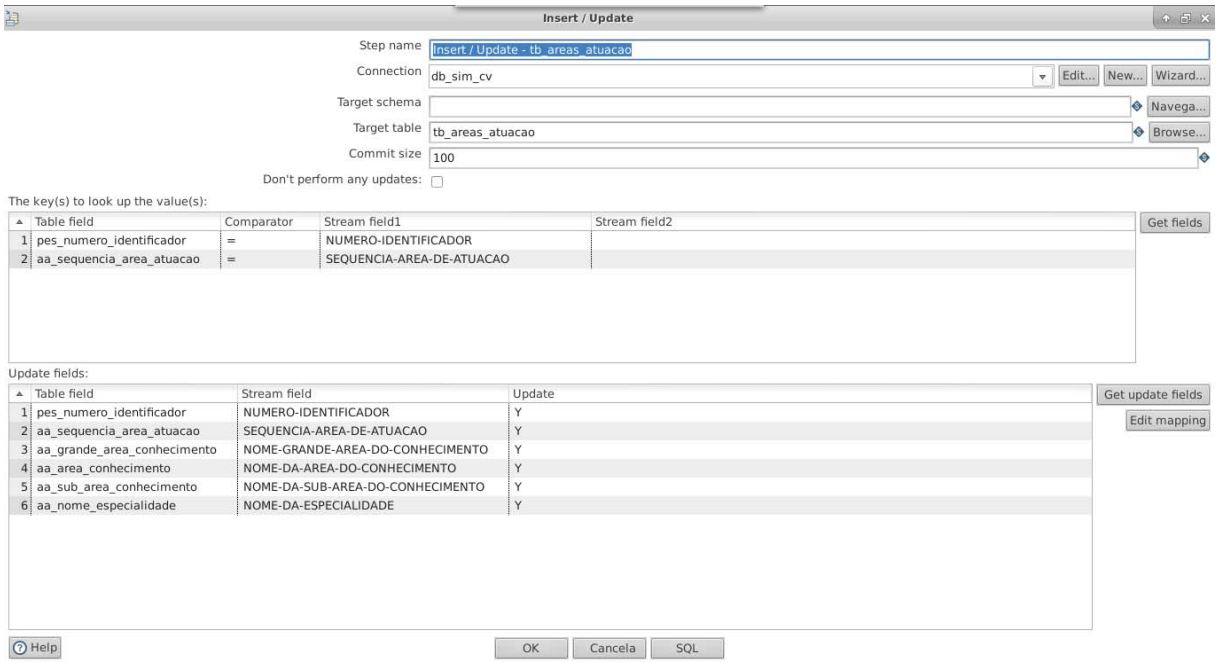


	Table field	Comparator	Stream field1	Stream field2
1	pes_numero_identificador	=	NUMERO-IDENTIFICADOR	
2	aa_sequencia_area_atuacao	=	SEQUENCIA-AREA-DE-ATUACAO	

	Table field	Stream field	Update
1	pes_numero_identificador	NUMERO-IDENTIFICADOR	Y
2	aa_sequencia_area_atuacao	SEQUENCIA-AREA-DE-ATUACAO	Y
3	aa_grande_area_conhecimento	NOME-GRANDE-AREA-DO-CONHECIMENTO	Y
4	aa_area_conhecimento	NOME-DA-AREA-DO-CONHECIMENTO	Y
5	aa_sub_area_conhecimento	NOME-DA-SUB-AREA-DO-CONHECIMENTO	Y
6	aa_nome_especialidade	NOME-DA-ESPECIALIDADE	Y

Figura 46: Correlação dos campos do XML exportado para os campos da tabela.

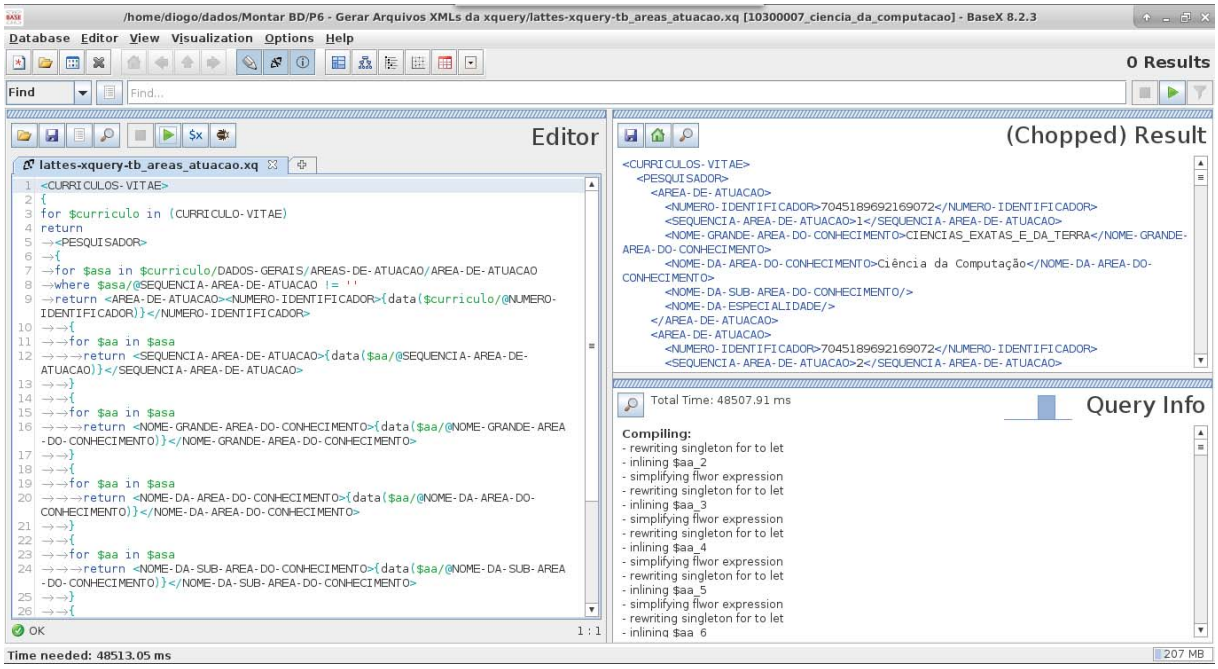


Figura 47: XQuery para exportação dos dados dos pesquisadores para XML específico, para então importar via ferramenta ETL.

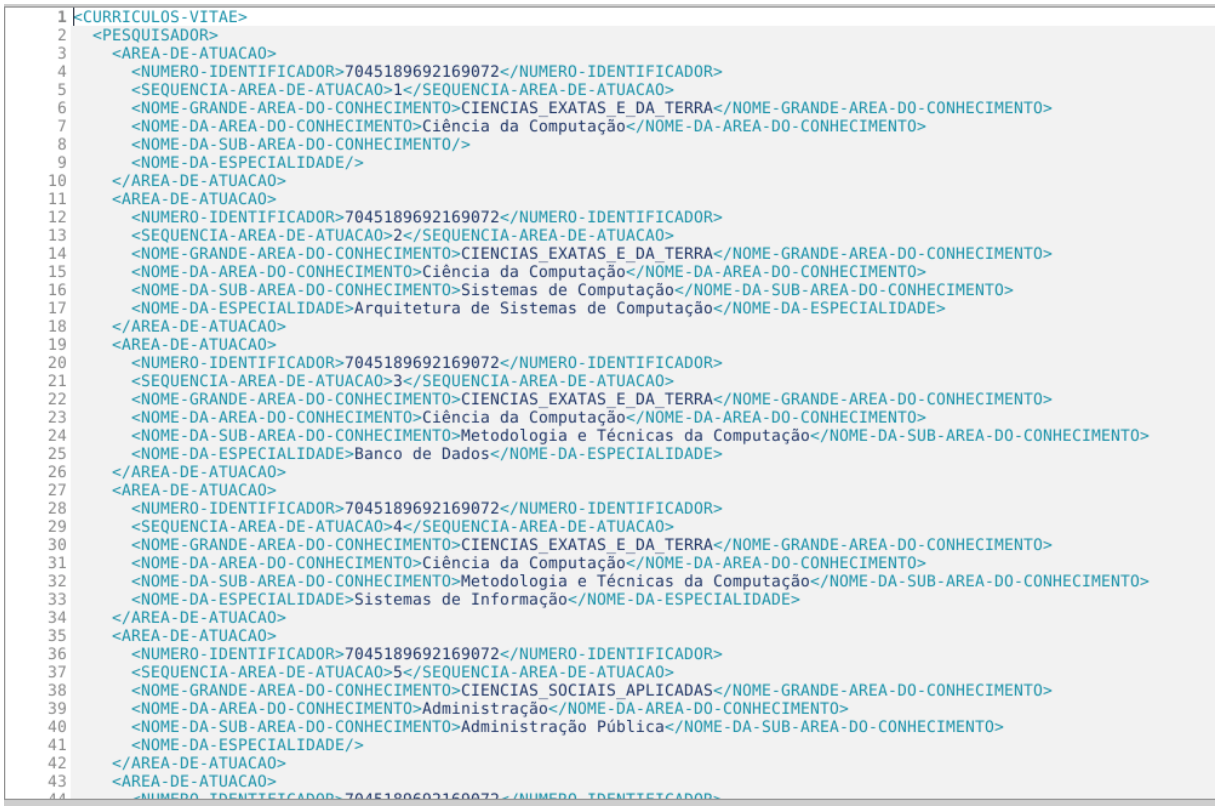


Figura 48: Exemplo de arquivo XML criado para carga nas tabelas da base de dados.

APÊNDICE C - CORRELAÇÃO DOS PESQUISADORES BOLSISTAS DE PRODUTIVIDADE DA CIÊNCIA DA COMPUTAÇÃO

Neste apêndice, através de imagens feitas do processo, demonstra-se os passos realizador para identificar quais currículos fazem parte dos bolsistas de produtividade da Ciência da Computação. Deste modo, os mesmos são correlacionados com os perfis já gerados posteriormente e utilizados nos experimentos.

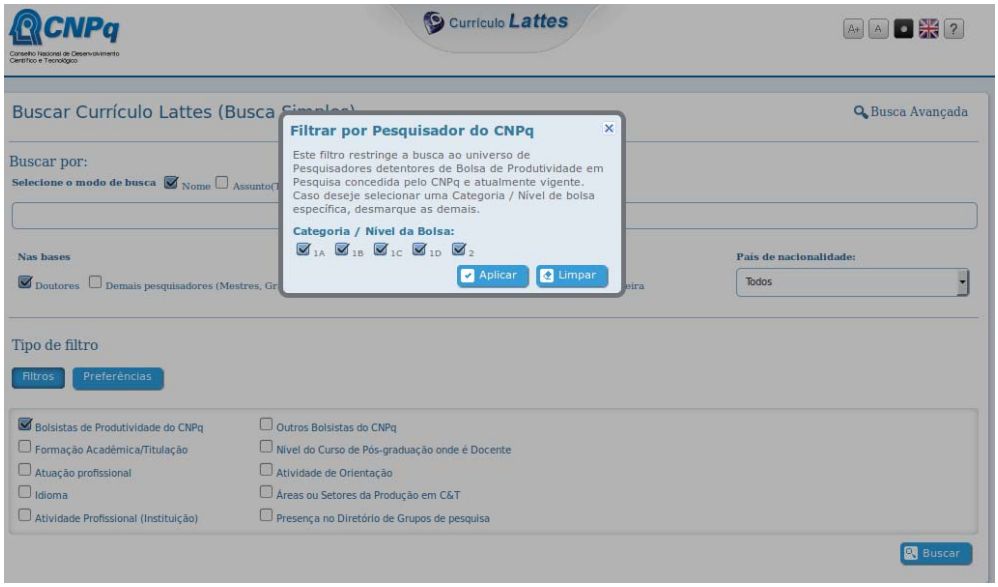


Figura 49: Plataforma Lattes, filtro de categoria/nível da bolsa utilizado para a busca dos bolsistas de produtividade.

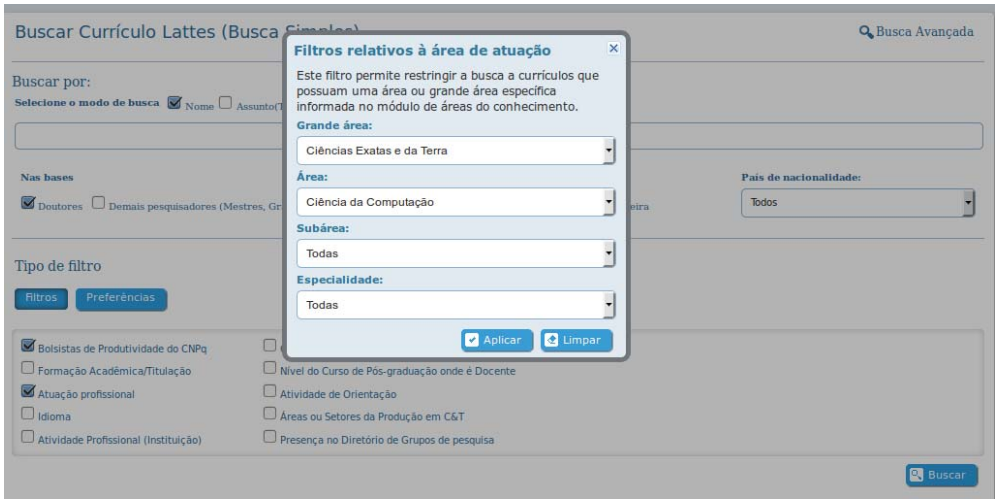


Figura 50: Plataforma Lattes, filtro de área utilizado para a busca dos bolsistas de produtividade.

Figura 51: Visão do código fonte do resultado da busca feita na Plataforma Lattes pelos bolsistas de produtividade da Ciência da Computação. No quadro vermelho, os identificadores.

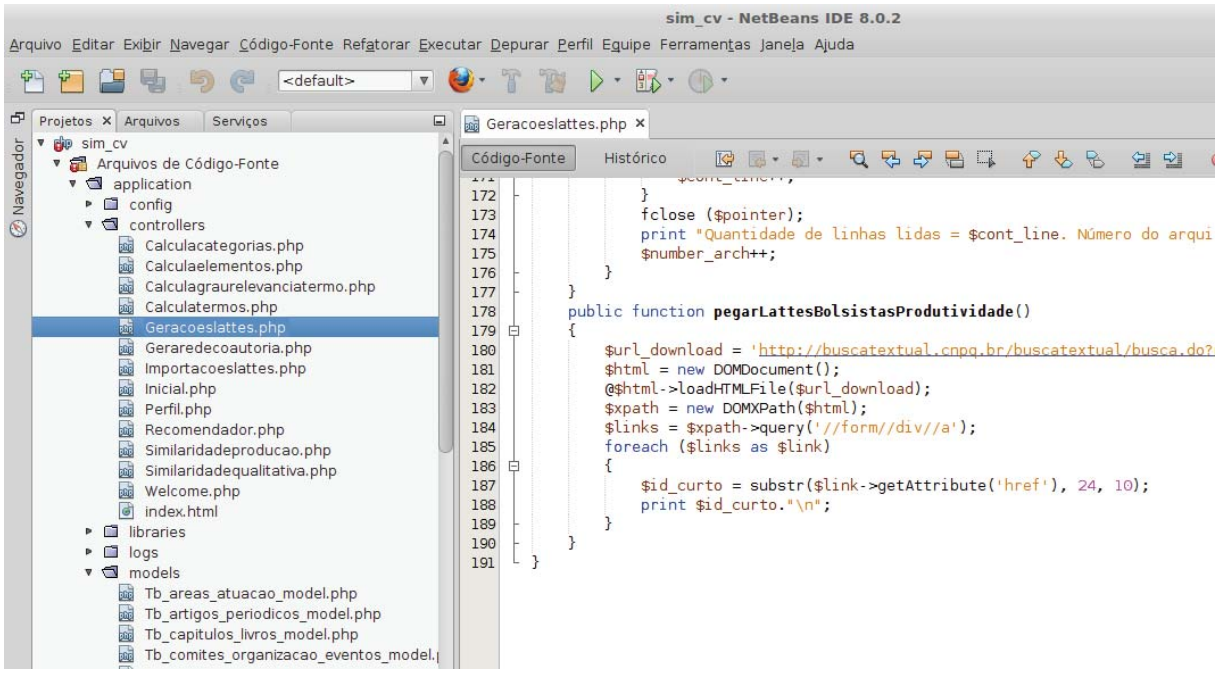


Figura 52: Código em PHP para extrair os identificadores dos bolsistas de produtividade retornados pela busca.

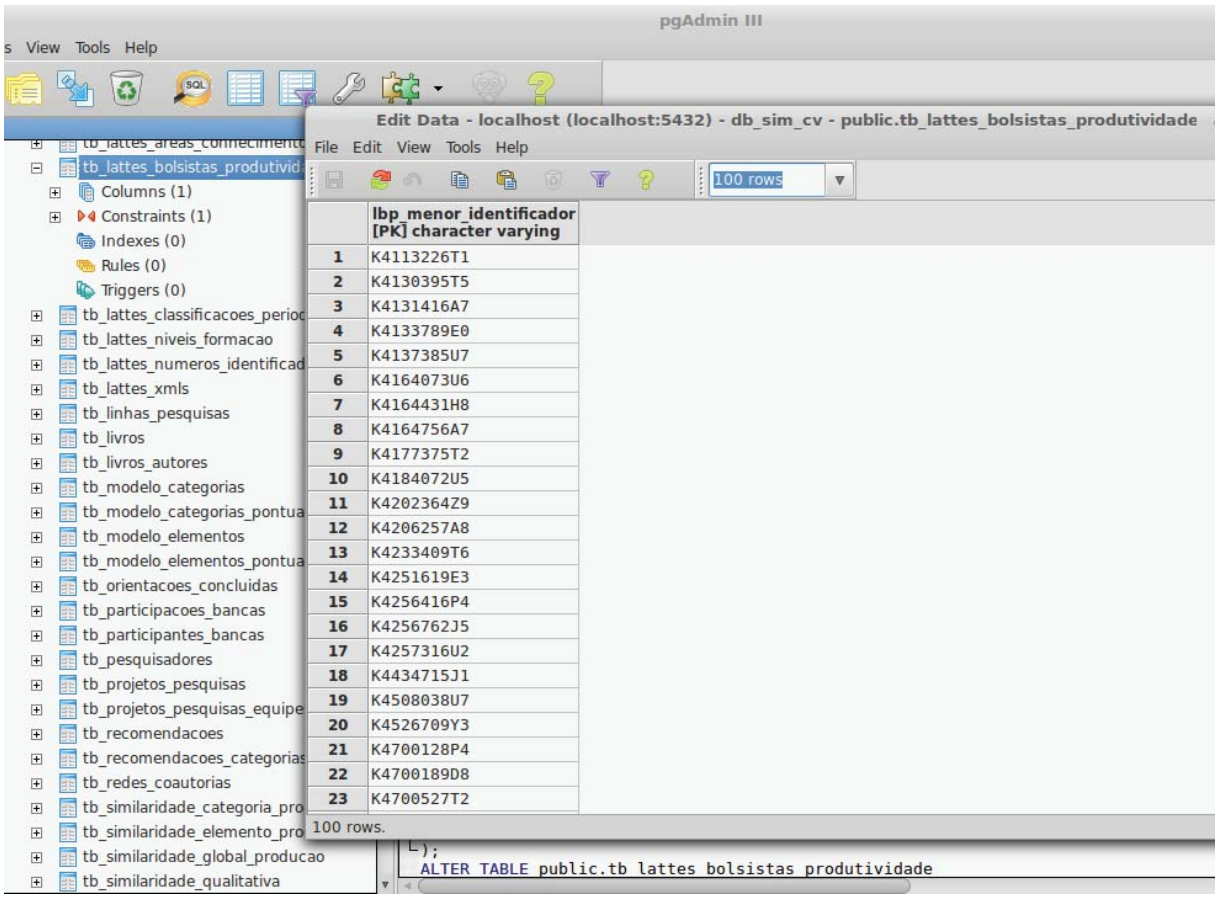


Figura 53:Tabela na base de dados com os identificadores dos bolsistas de produtividade da Ciência da Computação.