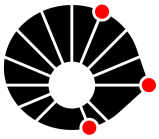
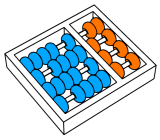
,





Anais do XVI Workshop de Teses, Dissertações e Trabalhos de Iniciação Científica do IC Unicamp

Technical Report - IC-21-06 - Relatório Técnico
December - 2021 - Dezembro

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Anais do XVI Workshop de Teses, Dissertações e Trabalhos de Iniciação Científica
Instituto de Computação - Unicamp

Apresentação

Este relatório técnico contém os resumos de 4 trabalhos cujos artigos foram autorizados a serem publicados no XVI Workshop de Teses, Dissertações e Trabalhos de Iniciação Científica (WTD)[1], do Instituto de Computação (IC) da Universidade Estadual de Campinas (Unicamp), edição 2021.

O XVI Workshop ocorreu entre os dias 15 e 16 de Dezembro de 2021 e contou com mais de 100 participantes, entre ouvintes ao vivo e apresentadores de trabalhos. O evento contemplou 4 artigos curtos, 16 lightning talks e 16 produções em vídeo. Aos alunos foi dada a possibilidade de escolher a forma de apresentação (lightning talk ou produção em video), bem como escolher se desejasse publicar ou não seu trabalho nos anais do evento. A publicação dos resumos sob forma de relatório técnico tem por objetivo divulgar os trabalhos em andamento e concluídos e registrar, de forma sucinta, o estado da arte da pesquisa do Instituto de Computação no ano de 2021.

Neste ano ocorreram 7 palestras em 2 dias de evento. A primeira, intitulada "Ciência Aplicada nas Empresas de Tecnologia", foi proferida pelo Dra. Juliana Medeiros, Cientista na área de Digital Security and Resilience na Microsoft. A segunda, intitulada "Estatística de Redes: Teoria, Métodos e Aplicações", foi proferida pelo Prof. Dr. André Fujita, do Departamento de Ciência da Computação, USP. A terceira intitulada "Fake News e Computação: Somos Parte do Problema e da Solução", foi proferida pelo Prof. Dr. Luiz Celso Gomes Jr, do Departamento de Informática, UTFPR.

A quarta palestra, "Paralelismo, Paralelismo, e Paralelismo... Como Programar um Supercomputador", foi proferida pelo Prof. Dr. Hervé Yviquel, do Instituto de Computação, Unicamp. A quinta palestra, "Stick to Your Goals to Reach Out for Self-adaptation", foi proferida pela Profa. Dra. Genaína Nunes Rodrigues do Departamento de Ciência da Computação da UNB.

A sexta palestra, "Utilzando a Web Semântica para Associar Conhecimentos dos Homens e das Máquinas", foi proferida pelo Dr. Marcos Da Silveira, Senior Researcher no Luxembourg Institute of Science and Technology, LIST. A última palestra, "Recuperação de Dados por Similaridade: Fundamentos, Operadores de Consulta e Indexação", foi proferida pelo Prof. Dr. Daniel dos Santos Kaster do Departamento de Computação, UEL.

Agradecemos aos alunos que participaram do evento, em particular àqueles que se dispuseram a apresentar seus trabalhos, seja oralmente ou em artigos curtos, bem como aos orientadores que os incentivaram a fazê-lo. Agradecemos, também, aos professores e pós doutorandos do IC que compuseram as bancas de avaliação dos trabalhos e aos colaboradores da secretaria que apoiaram a organização do evento. Agradecemos ao Professor Doutor Anderson de Rezende Rocha, diretor do IC, e a Professora Titular Cecília Mary Fischer Rubira, coordenadora da Pós-Graduação, pelo forte incentivo, apoio e patrocínio ao evento.

[1] https://www.ic.unicamp.br/wtd/2021/

Agradecemos às empresas Griaule, NeuralMind, ProFusion e ZenKlub, que engrandeceram o evento como patrocinadoras.

Finalmente, agradecemos imensamente aos alunos do programa de Pós-Graduação do IC que efetivamente organizaram o evento e que são coeditores deste relatório – André Gomes Regino, André Luiz do Canto Portela, Enio de Jesus Pontes Monteiro, Gustavo Caetano Borges, Helena de Almeida Maia, Lahis Almeida, Letícia da Silva Bomfim, Milene Elizabeth Rigolin, Víctor Jesús Sotelo Chico e Yan Prada Moro. A eles dedicamos o XVI Workshop de Teses, Dissertações e Trabalhos de Iniciação Científica do Instituto de Computação da Unicamp.

Prof. Julio Cesar Dos Reis
Profa. Esther Luna Colombini
Profa. Juliana Freitag Borin
Coordenadores do XVI WTD

Sumário

1 Programação

Apresentamos a programação e algumas estatísticas do XVI Workshop de Teses, Dissertações e Trabalhos de Iniciação Científica (WTD) do Instituto de Computação (IC) da Unicamp.

Nesta edição, tivemos 2 dias de evento. No primeiro dia (Figura 1) tivemos a abertura, 3 sessões de lightning talks, 6 palestras (2 de patrocinadores do evento).

	Aluno(a)	Trabalho	Orientador(a)	Banca
10:15-10:30	*Abertura (Prof. Julio Cesar dos Reis)*			
10:30-11:30	*Palestra: Ciência aplicada nas empresas de tecnologia (Chair: Prof. Julio C. dos Reis)*			
	Sessão 1 - Sistemas de Informação - 11:30 - 12:00 (Chair: Helena Maia)			
11:30-11:40	---	---	---	---
11:40-11:50	João Phillipe Cardenuto	Identificando Fábricas de Artigos Científicos Falsos com Análise de Proveniência	Anderson Rocha	Helena Maia
11:50-12:00	Rafael Soares Padilha	Aprendizado ativo na ordenação cronológica de eventos forenses	Anderson Rocha & Fernanda Andaló	Helena Maia
12:00-12:15	*Palestra: Montando times que acompanham as tendências tecnológicas - PROFUSION (Chair: Yan Prada)*			
	Almoço			
14:00-15:30	*Palestra: Estatística de Redes: teoria, métodos e aplicações (Chair: Prof. Julio C. dos Reis)*			
	Sessão 2 - Sistemas da Computação - 15:30 - 16:00 (Chair: Prof. Anderson Rocha)			
15:30-15:40	Tainá Turella Caetano dos Santos	Racismo Algorítmico em Redes Sociais	Islene Garcia	Prof. Anderson Rocha
15:40-15:50	---	---	---	---
15:50-16:00	Eduardo de Souza Gama	Video Streaming Analysis in Multi-tier Edge-Cloud Networks	Luiz Bittencourt	Prof. Anderson Rocha
16:00-16:20	*Coffee-break virtual*			
	Sessão 3 - Sistemas da Computação - 16:20 - 17:00 (Chair: André Portela)			
16:20-16:30	Jhonatan Cléto	Estendendo as funcionalidades do Task Bench para Benchmarking no OmpCluste	Hervé Yviquel	Juliana Freitag Borin
16:30-16:40	Carlos Alberto Astudillo Trujillo	Codificação na camada MAC para acesso aleatório massivo em redes 5G	Nelson Fonseca	Juliana Freitag Borin
16:40-16:50	Juliane Regina de Oliveira	Melhora da Qualidade dos Dados em Aplicações de Sensoriamento Remoto	Lucas Wanner	Sandro Rigo
16:50-17:00	Bruno Tojo da Silva	Funcionalidades do OMPC Bench para Benchmarking do OmpCluster	Hervé Yviquel	Sandro Rigo
17:00-18:30	*Palestra: Fake news e computação: somos parte do problema e da solução (Chair: Prof. Julio C. dos Reis)*			
18:30-18:45	*Palestra: Da UNICAMP ao Pentágono: tecnologia biométrica brasileira na vanguarda – e como fazer parte dela - GRIAULE (Chair: Yan Prada)*			
	Jantar			
19:30-20:30	*Palestra: Paralelismo, Paralelismo, e Paralelismo... Como Programar um Supercomputador !! (Chair: Prof. Julio C. dos Reis)*			

Figura 1: Programação do dia 1 do WTD

No segundo dia (Figura 2) tivemos 3 sessões de lightning talks e 5 palestras (2 de patrocinadores do evento), além do encerramento do WTD.

Dia 2 (16/12/2021) - Quinta feira				
	Aluno(a)	Trabalho	Orientador(a)	Banca
10:00-11:30	*Palestra: Stick to your goals to reach out for self-adaptation (Chair: Prof. Julio C. dos Reis)*			
	Sessão 4 - Teoria da Computação - 11:30 - 12:00 (Chair: Enio Monteiro)			
11:30-11:40	João Paulo Francisco da Silva	Teoria dos Jogos Algorítmica aplicada a alocação e precificação de recursos na Compu	Rafael Crivellari Saliba Schouery/L	Flávia Pisani
11:40-11:50	Tomás dos Santos Rodrigues e Silva	MDS Matrices for Cryptography	Ricardo Dahab	cmrubira
11:50-12:00	Fabio Akahoshi Collado	Inserting new information into NLP models to make them sensitive to legislative chang	Jacques Wainer	cmrubira
12:00-12:15	*Palestra: Sistemas Automatizado de Respostas Técnicas - Como a NeuralMind foi selecionada pelo SEBRAE - NEURALMIND (Chair: Enio Monteiro)*			
	Almoço			
14:30-15:30	*Palestra: Utilizando a Web Semântica para associar conhecimentos dos homens e das máquinas (Chair: Prof. Julio C. dos Reis)*			
	Sessão 5 - Sistemas de Informação - 15:30 - 16:00 (Chair: André Portela)			
15:30-15:40	Ilan Francisco da Silva	Segmentação Semi-Automática de Estruturas Torácicas em Exames de Tomografia	Alexandre Falcão	Helena Maia
15:40-15:50	Jose Italo da Costa Silva	Um estudo de caso para Sistema de Segurança Adaptativa: Implementação de um Firewall de Aplicação Auto-Adaptativo	Cecília Rubira	Breno de França
15:50-16:00	Flávia Érika Almeida Giló Azevedo	Inteligência Artificial Explicável Aplicada à Detecção Precoce da Doença de Alzheimer	Anderson Rocha	Arthur Valencio
16:00-16:30	*Coffee-break virtual*			
	Sessão 6 - Sistemas de Informação - 16:30 - 16:50 (Chair: Letícia)			
16:30-16:40	Gabriel Oliveira dos Santos	Desafios da audiodescrição automática em Português	Sandra Ávila/Esther Colombini	lehilton
16:40-16:50	Eryck Pedro da Silva	Análise dos Tópicos Mais Abordados em Disciplinas de Introdução à Programação em	Rodolfo Azevedo/Ricardo Caceffo	cmrubira
17:00-18:30	*Palestra: Recuperação de dados por similaridade: fundamentos, operadores de consulta e indexação (Chair: Enio Monteiro)*			
18:30-18:45	*Palestra: Produtos de dados e analytics no Zenklub - ZENKLUB (Chair: Yan Prada)*			
18:45-19:00	*Encerramento (Prof. Júlio Cesar dos Reis)*			

Figura 2: Programação do dia 2 do WTD

O evento contou também com sorteio de inúmeros brindes, fornecidos pelos patrocinadores do evento.

Os autores do WTD que se inscreveram na modalidade de vídeo compartilharam com o evento os vídeos relacionados as suas pesquisa, com duração de 10 a 15 minutos. Ao todo, 16 vídeos foram submetidos. Os 3 vídeos que tiveram mais visualizações e votações (votação via formulário do evento - peso 5; quantidade de likes no vídeo - peso 3; quantidade de visualizações - peso 2) foram premiados (Figura 3).

O evento foi finalizado com a sessão de brindes e premiações.

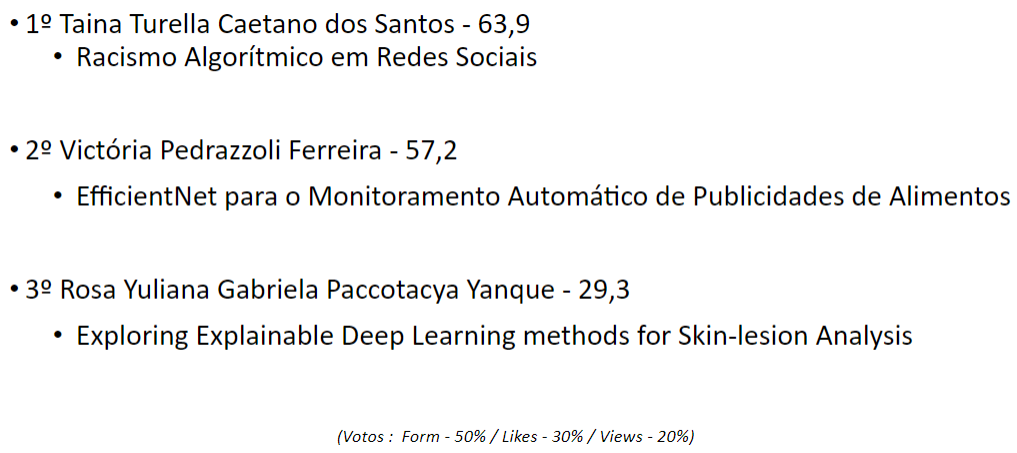


Figura 3: 3 Primeiros Colocados do Concurso de Vídeos do WTD

2 Estatísticas

Apresentamos as estatísticas colhidas em relação ao teor dos trabalhos apresentados durante o evento. A Figura 4 mostra a divisão dos trabalhos dentre os 3 tipos disponibilizados de submissão. A Figura 5 mostra a quantidade de alunos de graduação, mestrado e doutorado que apresentaram seus trabalhos. Por fim, a Figura 6 mostra as áreas de concentração dos trabalhos.

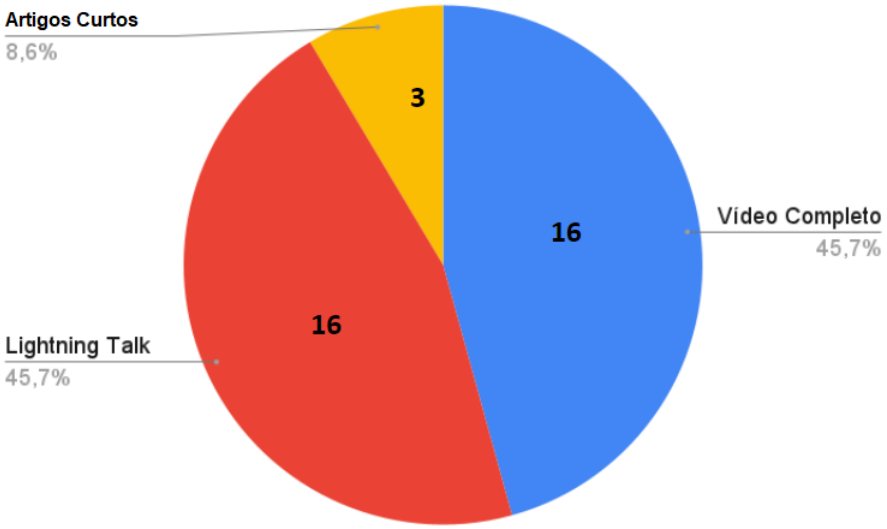


Figura 4: Tipos de Apresentações

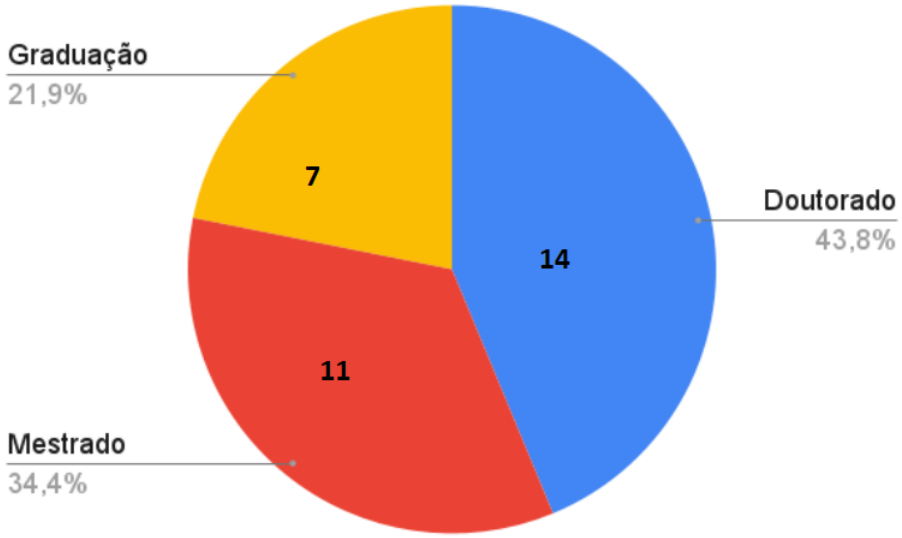


Figura 5: Titulação dos Alunos

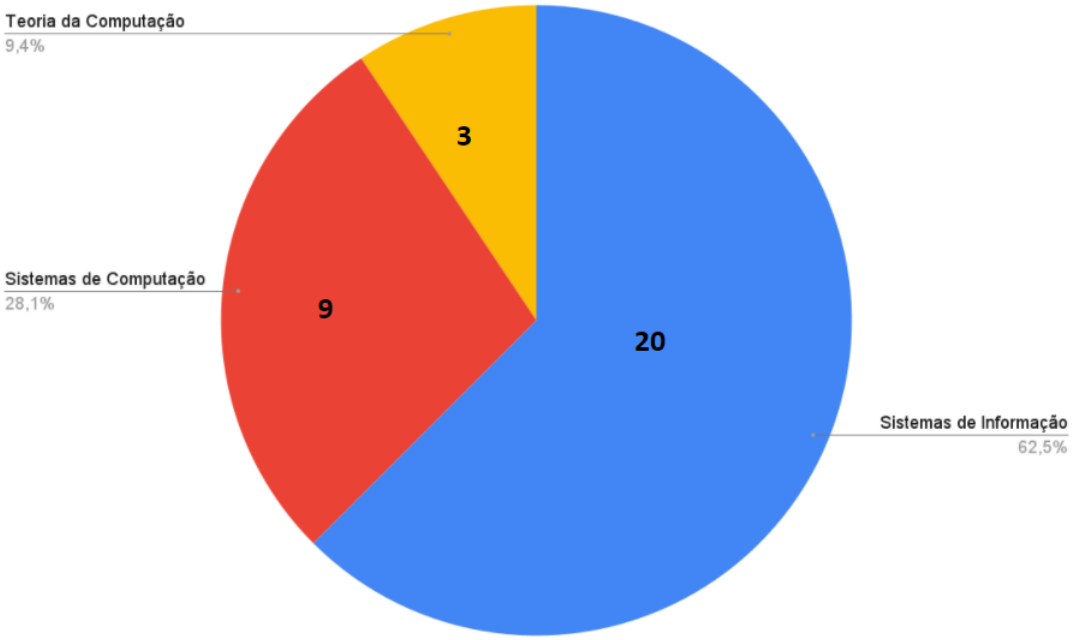


Figura 6: Áreas de Concentração

A Figura 7 apresenta a nuvem com termos mais frequentes dentre os trabalhos. Dentre esses termos, encontramos “analysis”, “learning” e “representation”.

Figura 7: Termos mais Frequentes

3 Artigos Curtos

Artificial Intelligence for the Discovery of Superconductors at Ambient Conditions

Camila M. Araújo[1], **Narcizo M. S. Neto**[2], **Anderson Rocha** [1]

[1]Institute of Computing - University of Campinas, Brazil (UNICAMP)

[2]Brazilian Synchrotrons Light Laboratory

c220742@dac.unicamp.br, narcizo.souza@lnls.br, arrocha@unicamp.br

Abstract. *A descoberta de um material supercondutor em condições ambientes seria revolucionária para todas as aplicações com eletricidade. Este projeto de pesquisa visa prospectar novos materiais teóricos com essa propriedade por meio da associação de simulações com dados experimentais. Pretendemos construir um modelo de previsão de estruturas cristalinas em condições extremas de pressão e temperatura com dados provenientes do Sirius, a fonte de luz Síncrotron da América Latina. Essas instalações experimentais oferecem as melhores condições de pesquisa do mundo para a prospecção de novos materiais supercondutores.*

Resumo. *The discovery of a superconducting material in ambient conditions would be revolutionary for all applications with electricity. This research project aims to prospect new theoretical materials with this property through the association of simulations with experimental data. We intend to build a prediction model of crystalline structures under extreme conditions of pressure and temperature with data from Sirius, the Sincrotron light source of Latin America. These experimental setups offer the best research facilities in the world for prospecting new superconducting materials.*

1. Introduction

Superconducting materials have zero resistance below a specific temperature (T_c). Therefore there is no loss of energy by conducting electricity. Another important property is the Meissner effect; the magnetic field is ejected from the material below the critical temperature. This phenomenon was discovered in 1911 when it was accidentally observed that mercury had zero resistance below 4 degrees Kelvin (K); hence it became a superconductor. Materials of this type are already widely used in magnetic resonance and maglev trains, which work with magnetic levitation.

Recent discoveries indicate that super-hybrid materials subjected to very high pressures can exhibit superconducting properties [Bi et al. 2018]. This is due to characteristics of the hydrogen atom that generate conditions favorable to showing superconductivity at high temperatures. For example, in 2015, a group of researchers at the *Max Planck Institute for Chemistry* found that hydrogen sulfide becomes a superconductor at 203K when subjected to a pressure of 150GPa; under these conditions, the superconductor H_3S is formed [Drozdov et al. 2015]. In practice, several new materials with high T_c were first discovered by quantum simulations, such as H_3S and LaH_{10}, [Liu et al. 2017], and

then experimentally proven [Somayazulu et al. 2019]. Thus, the theoretical discovery of new superconductors can be divided into two steps: predicting new crystal structures and calculating T_c. The prediction of new crystal structures is a global optimization problem. From a set of atoms, we seek to obtain the spatial arrangement whose structure energy is minimal. The lowest energy structure is the most stable and most likely to form in the real world. CALYPSO (*Crystal structure AnaLYsis by Particle Swarm Optimization*) [Wang et al. 2012] is a software package developed for this purpose. LaH_{10}, for example, was discovered using this tool.

Materials Sciences research is increasingly evolving towards working with inverse design problems. The aim is no longer to find the structure with the lowest energy but the structure that maximizes a desired property, such as the superconductivity [Schleder et al. 2019]. These techniques are state of the art in this area, better known as Materials Informatics. Currently, the search for new theoretical materials comes from researchers' insights, such as the choice of atoms and input conditions for tools like CALYPSO to predict promising crystal structures. Computer Science can improve this process' efficiency by adding intelligence to this step. Modern science is moving towards a new data-driven paradigm. It seeks to obtain physical insights from data in contrast to the traditional approach of fitting the data to the theory [Schleder et al. 2019]. In this context, Artificial Intelligence (AI) tools play a crucial role. The superconductivity problem benefits from AI techniques to build knowledge as there is no complete physical theory for the phenomenon. AI tools allow us to combine experimental data analysis, *ab initio* simulations and physical insights in the search for new materials with the desired properties.

In 2020, a group of researchers obtained the first superconductor at room temperature. A compound formed by hydrogen, sulfur, and carbon presented T_C of 293K (15 ºC) at a pressure of 267 GPa [Snider et al. 2020]. This remarkable discovery is yet another indication that obtaining a superconducting material at room temperature is just a matter of time. However, it was not possible to determine the chemical formula or crystal structure of the compound, the light quantum nature of hydrogen limits this experimental determination by X-ray scattering. This is not a trivial problem, and a possible solution is to combine experimental and theoretical approaches that complement each other. Connecting *ab initio* simulations to experimental research allows for complete predictions, as the results obtained independently validate each other. Thus, it is possible to obtain experimental proof of the superconductivity phenomenon, the corresponding molecular formula, and the crystal structure. The Brazilian Synchrotron Light Laboratory (LNLS), a partner in this project, has a unique infrastructure to allow this type of joint investigation. The relevance of computational solutions and AI tools is enormous in this type of study, and their potentials have just begun to be explored.

2. Proposal

We will develop, implement and test algorithms to search for theoretical new superconductor materials in ambient conditions.

2.1. Research Goals

The ultimate goal of this research is to discover a new superconducting material at ambient temperature and pressure. For this purpose, we aim to develop an intelligent searching

tool for discovering materials based on optimizing the desired property, superconductivity. Our secondary goals include, but are not limited to:

1. Collect a data set from Sirius experiments with a combination of crystalline structures formed, samples' chemical elements, and samples' pressure and temperature;
2. train a Machine Learning (ML) model to learn what crystalline structure will be formed when submitting a sample to extremely high pressures and temperatures;
3. automate the obtainment of T_c for a predicted structure from quantum mechanics simulations;
4. develop an intelligent algorithm to prospect new superconducting materials based on the combination of elements from the periodic table and extreme pressure and temperature conditions.

2.2. Prior Work

In previous work at the Sirius' beamline EMA, the student developed a software solution to automate quantum simulations for T_c [Araújo 2020]. The theoretical calculation of T_c is essential to predicting superconductors, allowing the analysis of both known and theoretically new structures seeking high T_c. The package available to execute this task, Quantum ESPRESSO (QE) [P Giannozzi], needed several complex manual configurations and different steps to calculate the property. In addition, it was subject to numerous human typing errors and inconsistencies. Because of these issues, the preparation time for simulations was long and could cost days for only one compound. The complete automation of the simulations needs only one input file: a crystal structure file (.cif), the international standard representation of materials structures. The software automation decreased days of work to a few seconds. With this tool, structures obtained from Sirius' experimental data can be directly used as input for the program that provides various essential properties, including T_c. It represented a step towards a computerized searching tool for discovering materials based on optimizing desired properties.

3. Proposed Methodology

3.1. Data acquisition and structures' prediction

In order to discover new materials, we need to predict how a set of atoms will organize to form a crystalline structure, which can be described in terms of its unit cell geometry, the smallest repeating unit having the full symmetry of the crystal. A unit cell has six lattice parameters, three distances, and three angles. According to its unit cell, all existing materials can be classified into 14 three-dimensional Bravais lattices and 230 space groups according to their symmetry. There are already successful structure prediction ML models for ambient conditions. One example is CRYSPNet [Liang et al. 2020]. This model predicts the bravais lattice, space group, and lattice parameters of a structure using as input the chemical composition and predictors that aggregate properties of the elements constituting the compound. A series of neural network models trained with more than 100.000 structures from the Inorganic Crystal Structure Database (ICSD) was used, and the obtained results were excellent. Another important and recent work is MlatticeABC [Li et al. 2021], a random forest ML model to predict lattice unit cell lengths that showed significant improvement in angle prediction as well. These works showed

promising paths; we intend to explore these methodologies for structure prediction in extreme conditions, which are prone to forming superconductors. This will be done with a data-driven approach, using a dataset obtained from real structures subjects to extreme conditions at Sirius.

Sirius is one of the first fourth-generation synchrotron light sources in the world, located in Campinas, SP, as part of The Brazilian Synchrotron Light Laboratory (LNLS). This infrastructure allows the realization of several cutting-edge researches on the microscopic structure of matter. One of these is the project "A look with X-rays into superconductivity,"led by LNLS researcher Narcizo Marques de Souza Neto, who will be the main contributor to this project. The experiments will be conducted at EMA beamline (Extreme Methods of Analyzes), which can perform experiments at pressures near the Earth's center. EMA is capable of generating a pressure map. Each sample, of a determined compound and approximately 100 square microns, is subject to a high pressure and analyzed in the beamline. The distribution of pressure is no-homogeneous across the sample, therefore a gradient is formed, which originates hundreds or thousands of gradient pressure points, corresponding each to a different structure. Adding to this the possibility of varying the pressure in one experiment and testing numerous different samples, the data generated grows exponentially. This high throughout data will be used to feed a ML model to predict crystalline structures formed in the function of pressure. This model will allow the prospect of new theoretical superconductor materials with low computational cost compared to the existing solutions as CALYPSO. Because of this, a scanning algorithm to execute an oriented search for superconductors will be computationally doable.

3.2. Materials inverse design

Our data-driven approach (3.1) will be fundamental to create new theoretical high-pressure structures, using as input the chemical composition and extreme ambient conditions. We aim to integrate the previous solutions into an intelligent searching tool to prospect new materials. With the previous work of DFT Automation (2.2), we can directly obtain the T_c of each generated structure and use these values as heuristics for the searching algorithm to maximize the desired superconductivity temperature. The partnership with Sirius specialists will be rich to filter non-physical results and refine the optimization process until convergence or elimination of unpromising combinations. We will also be able to add explainability to the obtained results.

One approach to the inverse design of new materials is to search for a global minimum with local optimization techniques. The most popular methods used are genetic algorithms and basin hopping optimization [Schleder et al. 2019]. The basin hopping algorithm starts with a randomly chosen and deformed structure. The structure is brought to a local minimum by relaxation. If the new energy minimum is lower than the previous one, the movement will be accepted. The cycle repeats itself, efficiently probing the reciprocal space (possible geometries for the structure). With evolutionary algorithms, the search space is constrained by choosing surviving structures that maximize the desired property, its intrinsic parallelism also allows the search in multiple directions.

4. Expected Results

We expect to find a set of viable theoretical materials, each consisting of a chemical formula, crystalline structure, and expected high values of T_c. The Sirius research group will

further study this set of superconducting candidates, and the most promising structures will be select to be synthesized and tested. We will also dispose to the scientific community a new method to predict crystalline structures as a function of pressure: an ML model based on results from Sirius experiments. It has the advantage of being significantly less computationally costly than the existing algorithms based on quantum simulations. This work will also trace new paths to understand the phenomenon of superconductivity and serve as a foundation to the rising field of Materials Informatics.

Referências

Araújo, C. M. (2020). Automation of dft calculations for superconductors. *30thRAU Abstract Book*, 2(1):53.

Bi, T., Zarifi, N., Terpstra, T., and Zurek, E. (2018). The search for superconductivity in high pressure hydrides. *arXiv preprint arXiv:1806.00163*.

Drozdov, A., Eremets, M., Troyan, I., Ksenofontov, V., and Shylin, S. I. (2015). Conventional superconductivity at 203 kelvin at high pressures in the sulfur hydride system. *Nature*, 525(7567):73–76.

Li, Y., Yang, W., Dong, R., and Hu, J. (2021). Mlatticeabc: generic lattice constant prediction of crystal materials using machine learning. *ACS omega*, 6(17):11585–11594.

Liang, H., Stanev, V., Kusne, A. G., and Takeuchi, I. (2020). Cryspnet: Crystal structure predictions via neural networks. *Phys. Rev. Materials*, 4:123802.

Liu, H., Naumov, I. I., Hoffmann, R., Ashcroft, N., and Hemley, R. J. (2017). Potential high-tc superconducting lanthanum and yttrium hydrides at high pressure. *Proceedings of the National Academy of Sciences*, 114(27):6990–6995.

P Giannozzi, e. a.

Schleder, G. R., Padilha, A. C., Acosta, C. M., Costa, M., and Fazzio, A. (2019). From dft to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials*, 2(3):032001.

Snider, E., Dasenbrock-Gammon, N., McBride, R., Debessai, M., Vindana, H., Vencatasamy, K., Lawler, K. V., Salamat, A., and Dias, R. P. (2020). Room-temperature superconductivity in a carbonaceous sulfur hydride. *Nature*, 586(7829):373–377.

Somayazulu, M., Ahart, M., Mishra, A. K., Geballe, Z. M., Baldini, M., Meng, Y., Struzhkin, V. V., and Hemley, R. J. (2019). Evidence for superconductivity above 260 k in lanthanum superhydride at megabar pressures. *Physical review letters*, 122(2):027001.

Wang, Y., Lv, J., Zhu, L., and Ma, Y. (2012). Calypso: A method for crystal structure prediction. *Computer Physics Communications*, 183(10):2063–2070.

Redes Sociais - Filtros de relevância, enviesamento algorítmico e ações afirmativas

Tainá Turella C. dos Santos
Prof. Dra. Islene Calciolari Garcia

[1]Instituto Computação – Universidade Estadual de Campinas (UNICAMP)
Caixa Postal 13083-852 – São Paulo – SP – Brazil

t187251@dac.unicamp.br, islene@unicamp.br

Abstract. *Social networks, such as Facebook, Twitter, Instagram, etc., is a huge mirror of the existing society. Through them, individuals can express themselves while also hearing others, but those platforms are also a source of discrimination, not just due to its users, but due to the recommendation algorithm behind it. Some users had already noticed that, at least on Instagram, there is a palpable difference in posts delivery and range between white and black content producers. This paper aims to provide a guide on how this difference can be measured, with the creation and analysis of an open dataset, and also proposes a new manner to address recommendations to enhance equality between producers with "affirmative actions"embedded in the code.*

Resumo. *Redes sociais como Facebook, Twitter, Instagram, etc., são um grande espelho da sociedade existente. A partir delas, indivíduos podem se expressar e ouvir demais pessoas, entretanto estas plataformas também se tornam uma fonte de discriminação, não apenas por conta de ações diretas de seus usuários, mas também devido aos algoritmos de recomendação por trás delas. Alguns usuários notaram, tendo como exemplo o Instagram, que existe uma diferença palpável entre as métricas de entrega e alcance de produtores de conteúdo negros e brancos. Este artigo tem como objetivo prover um guia de como medir e analisar tal diferença em um dataset aberto, mas também se propõe a aumentar a igualdade entre produtores de conteúdo com o uso de "ações afirmativas"embutidas no próprio código.*

1. Contexto

O fenômeno das redes sociais é algo espantoso, pois quem poderia imaginar que qualquer pessoa poderia ter a oportunidade de realizar novas conexões ou compartilhar ideias em escala global? Os filmes e desenhos, de décadas anteriores, retratavam um futuro com carros voadores, tênis que se amarrariam sozinhos, etc., mas que não foram capazes de prever, não em mesma escala, as redes sociais.

Plataformas como *Facebook*, *Twitter*, *Instagram* e similares atualmente representam uma grande parte do tráfego na internet, e isso mostra o quanto tais ferramentas já se tornaram parte do cotidiano de diversas pessoas, não só como uma fonte de distração, mas também como ferramenta de trabalho e portanto fonte de renda de muitos indivíduos. Atualmente grande parte da divulgação de conteúdo escoa pelas redes sociais e é necessária uma análise de como os algoritmos, por trás dessas plataformas, aprendem e reproduzem

certos padrões, principalmente quando prejudicam uma parcela considerável de usuários por conta de discriminações diversas como;gênero, credo ou raça.

As redes sociais possuem características diversas, mas um dos pontos que mais atrai seus membros é a capacidade das mesmas em apresentar conteúdos filtrados e categorizados de forma a atender melhor os gostos pessoais de cada um. Mas algumas perguntas ficam no ar; "Será que no meio desse processo de recomendação algumas escolhas feitas pelas IAs (Inteligências Artificiais) não foram baseadas em algum viés retrógrado da sociedade?", "Será que alguns conteúdos são preteridos a outros?", "Será que alguns produtores de conteúdo possuem menos alcance que os demais por conta da política de recomendação?".

Existem evidências [Levi K. 2020] de que os algoritmos de relevância e entrega de redes sociais famosas possuem um enviesamento racial, mas em redes de código aberto como *PixelFed*, *Barinsta*, *Karma*, *Mastodon*, etc., até por não serem tão utilizadas como plataformas de divulgação de conteúdo em massa, tais relatos são escassos. Entretanto, será que é possível sair de um plano de achismos e criar uma base estruturada que prove a influência de questões raciais em tais algoritmos?

É um fato, conhecido pelas pessoas que estudam a área de *Machine Learning*, que os algoritmos podem vir a tomar decisões baseadas em estigmas sociais, que podem levar a condutas que são classificadas como misóginas ou racistas, por exemplo. Existem trabalhos nesta área que buscam compreender como, uma determinada base de dados, "ensina"a máquina a ter certas atitudes extremamente condenáveis e em suma se chega a conclusão de que os algoritmos por si só não são racistas ou machistas, mas os dados de treinamento muitas vezes o são.

1.1. Revisão Bibliográfica

Em diversas áreas do conhecimento, pesquisas sobre como o racismo estrutural têm sido realizadas, e a ideia de racismo algorítmico tem sido discutida cada vez mais no meio tecnológico. A ilusão de que artefatos computacionais são neutros é algo que precisa ser superado para que o problema realmente possa ser solucionado. Pesquisas como a de Safiya Noble [Noble 2018] e de Tarcízio Silva, dentre muitos outros [Silva 2020], mostram o quanto o estudo e a aplicação de conceitos éticos em computação é necessário.

Estudos sobre enviesamento racial na área de *Machine Learning* têm despontado cada vez mais no meio acadêmico. Grande parte desses estudos tem seu foco em compreender como *datasets* de imagem podem gerar um algoritmo falho e muitas vezes racista, assim como apontado nos trabalhos de Joy Buolamwini [Raji et al. 2020, Buolamwini and Gebru 2018]. Suas descobertas sobre a ineficiência do reconhecimento de faces negras por diversos produtos comumente utilizados no mercado são uma prova de que os algoritmos por si só não são enviesados, mas a forma pela qual são treinados leva a tomadas de decisões minimamente preocupantes quando se busca uma equidade social.

Se é feita uma análise mais ampla, sobre os efeitos da aplicação de IA em diversas frentes da vida em sociedade, têm-se em Cathy O'Neil uma discussão profunda sobre como tais algoritmos podem ser utilizados de forma a prejudicar minorias ou pessoas em condições de vulnerabilidade social. Em seu livro, *Weapons of Math Destruction* [O'Neil 2016], a autora dá exemplos práticos nos Estados Unidos onde tal tecnologia

é utilizada com a intenção de fazer com que a sociedade seja mais segura, mas acaba gerando um *feedback* que tende a prejudicar de forma clara as populações latino-africanas do país.

Por conta de casos amplamente divulgados [Vincent 2016, BBC 2015], sabe-se que os algoritmos por trás do aprendizado das IAs podem se tornar racistas por sua dependência de dados que são fornecidos por uma sociedade que, infelizmente, possui certos vieses e certos preconceitos profundamente enraizados no pensamento coletivo. Existem pesquisas [de Gibert et al. 2018] cujo foco é a identificação de discursos de ódio na internet e sabe-se que, mesmo em redes sociais menos divulgadas [Robertson 2019] são aplicadas medidas para a contenção deste tipo de discurso explícito. Mas algo que é necessário manter em mente é que o racismo, por exemplo, muitas vezes é declarado de forma sutil.

Pensando mais no quesito de redes sociais, pode-se encontrar uma base de material de estudo em Tufecki. No livro *Twitter and Tear Gas* [Tufekci 2017] encontra-se um relato de como as plataformas de comunicação social se tornaram uma fonte de informação, mas também é possível perceber que a existência de filtros em algumas destas redes oculta certos acontecimentos que podem ter uma relevância política e social muito maiores do que qualquer *"meme"* publicado por um perfil qualquer.

Sabe-se que a questão de algoritmos de recomendação, relevância e a entrega/alcance de *posts* têm sido notada por alguns usuários ativistas negros/negras, mas o que ainda é escasso no meio é uma discussão acadêmica, técnica e social do assunto. O diferencial deste projeto será a compreensão desses algoritmos e a proposição de uma forma de minimizar seus impactos negativos sobre a população negra.

2. Objetivo

Tendo em mente os pontos levantados durante a contextualização deste texto, redes sociais e viés algorítmico, propõe-se um estudo da união destes. A ideia deste projeto é criar uma base argumentativa, em cima da coleta de dados sobre alcance de *posts* no *Instagram* de perfis diversificados, de forma a poder realizar uma análise se as métricas foram impactadas por algum tipo de viés racial. Para que, a partir dessa base sólida, seja possível propor um algoritmo que propositalmente garanta que entre os *posts* ou perfis mais relevantes, também possam ser encontradas produções de pessoas negras.

Esta proposta de algoritmo de recomendação será baseada no algoritmo PageRank [Xing and Ghorbani 2004], porém algumas modificações serão propostas para que exista uma possibilidade de aumentar a visibilidade e o rankeamento de perfis e posts feitos por pessoas negras. Propõe-se, basicamente, um algoritmo que possua o sistema de ações afirmativas embutido em seu código.

3. Método

Parece válido separar a metodologia envolvida nos processos deste trabalho em duas etapas, a primeira para a coleta e análise dos dados obtidos e uma segunda para o próprio desenvolvimento do algoritmo de recomendação que possuirá as ações afirmativas.

3.1. Primeira etapa

Sabendo que os dados que esta pesquisa se propõe a analisar são dados oriundos do Instagram, uma rede social de código fechado que possui aproximadamente 1,22 bilhões de usuários ativos, pode-se imaginar que existem diversos entraves para ter acesso, mesmo que de forma anonimizada aos dados dos usuários da rede. Desta forma, foi definido que tal busca por dados será feita via contato direto com influencers via "Direct Messages" do Instagram e que também será feita uma coleta através de um formulário [1] que contém um Termo de consentimento livre e esclarecido.

Os dados, que serão encaminhados pelos voluntários da pesquisa, serão obtidos a partir de uma ferramenta de Web Scrapping [2] rodando sobre a página do Facebook Creator Studio de acesso exclusivo a pessoa responsável pelo perfil, do tipo Business, no Instagram.

Em cima dos dados coletados pretende-se fazer uma análise comparativa de perfis afins, cuja única diferença seria um fator racial, para que se possa compreender de forma quantitativa como o viés racial afeta as pessoas negras e a distribuição de seu conteúdo em redes sociais de código fechado. é parte do planejamento realizar o mesmo tipo de análise sobre dados de redes sociais de código aberto, de forma a compreender como cada uma destas "comunidades" se comporta.

3.2. Segunda Etapa

Nesta segunda etapa acredita-se que será necessário treinar alguns modelos com auxílio de técnicas de Machine Learning para que seja possível encontrar os valores corretos dos pesos que devem ser atribuídos às arestas do algoritmo PageRank modificado, um exemplo pode ser encontrado na Figura 1. Em cima destes modelos serão repassados os dados coletados de forma a testar se as ações afirmativas estão causando o efeito esperado, fazendo com que perfis de pessoas negras tenham uma maior visibilidade no processo de recomendação.

Após uma fase inicial de testes será feito o acoplamento do algoritmo desenvolvido a rede social de código aberto PixelFed. Desta forma será possível deixar o algoritmo aberto para que mais estudos sejam feitos sobre ele, além de causar um impacto positivo na comunidade que apoia o movimento Software Livre.

4. Conclusão

Com base no que foi exposto até o momento, acredita-se que existem indícios para se afirmar que o Racismo Algorítmico está presente nas redes sociais, e que este possui um impacto extremamente negativo sobre as pessoas negras da área de produção de conteúdo. Espera-se que com mais estudos seja possível trazer provas mais concretas que suportem tal afirmação.

O tema Racismo Algorítmico se tornou cada vez mais popular, mas ainda existem poucas produções acadêmicas, vindas de pesquisadores da área de Computação, que tenham como objetivo não só expor mas também solucionar este tipo de discriminação. É

[1] O formulário em questão estará disponível em breve no site `www.tainaturella.com`.

[2] Um vídeo, explicando como a ferramenta de We Scrapping precisa ser configurada, será liberado para acesso no site `www.tainaturella.com`.

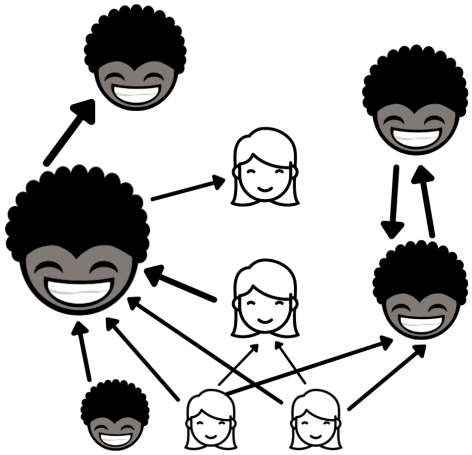


Figura 1. PageRank modificado para que certas arestas possuam mais relevância que outras dada sua fonte

necessário falar, ouvir, pesquisar e tomar atitudes sobre esta mais nova forma de racismo, para que em algum momento possamos realmente alcançar uma sociedade que oferece as mesmas condições para todas as pessoas que a compõe.

Referências

BBC (2015). Google apologises for photos app's racist blunder. Last accessed April 28th 2021.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, New York, NY, USA.

de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate speech dataset from a white supremacy forum.

Levi K., F. (2020). Racismo algorítmico não é apenas sobre engajamento em redes sociais. Last accessed May 05th 2021.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA.

Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 145â151. Association for Computing Machinery, New York, NY, USA.

Robertson, A. (2019). How the biggest decentralized social network is dealing with its nazi problem. Last accessed April 28th 2021.

Silva, T. (2020). *Comunidades, Algoritmos e Ativismos Digitais: Olhares Afrodiaspóricos*. LiteraRUA, São Paulo.

Tufekci, Z. (2017). *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press, USA.

Vincent, J. (2016). Twitter taught microsoft's ai chatbot to be a racist asshole in less than a day. Last accessed April 28th 2021.

Xing, W. and Ghorbani, A. (2004). Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pages 305–314.

Análise dos Tópicos Mais Abordados em Disciplinas de Introdução à Programação em Nível Superior

Eryck Pedro da Silva[1]*, **Ricardo Edgard Caceffo**[1], **Rodolfo Jardim de Azevedo**[1]

[1]Instituto de Computação – Universidade Estadual de Campinas

{eryck.silva,caceffo,rodolfo}@ic.unicamp.br

Abstract. *This work, part of a Ph.D thesis, conducted an analysis of the syllabi of 150 introductory programming courses (CS1), gathered from 61 Brazilian Federal Universities, in order to identify the most covered topics. The results are composed of 12 topics regarding structured programming, which are also compared with listings that are present in similar work, as well as the ones covered by UNICAMP's MC102 subject. The main objective was to report, in a thorough manner, the most covered CS1 topics in Brazil, aiming to help justify future interventions regarding the teaching and learning of these topics.*

Resumo. *Este trabalho, representando parte de uma tese de doutorado, realizou uma análise de ementas e conteúdos programáticos de 150 disciplinas de introdução à programação (CS1), contemplando 61 Universidades Federais brasileiras, para identificar os tópicos mais abordados. Os resultados são compostos por 12 tópicos relacionados à programação estruturada, também comparados com listagens presentes em trabalhos semelhantes, bem como a disciplina de MC102 da UNICAMP. O objetivo principal foi reportar, de forma abrangente, os tópicos mais abordados em disciplinas de CS1 no Brasil, objetivando apoiar a justificativa de futuras intervenções no ensino e aprendizagem desses conteúdos.*

1. Introdução

A utilização da computação enquanto ferramenta continua em expansão nas mais diversas áreas, seja para tarefas comuns, produção de conhecimento ou no mercado de trabalho. No ensino superior, os cursos da áreas de ciências exatas e engenharias são os que costumam oferecer disciplinas de computação, em especial, as em nível introdutório que ensinam programação de computadores [Nascimento 2018]. De forma que os nomes dessas disciplinas introdutórias costumam variar bastante, inclusive numa mesma Instituição de Ensino Superior (IES), este trabalho optou por utilizar o termo *CS1* (*Computer Science 1*) em referência ao currículo desenvolvido pela *Association for Computing Machinery* (ACM) em 1978 [Austing et al. 1979], que, seguido por *CS2*, correspondem às duas primeiras disciplinas de programação cursadas por um aluno de graduação na área de computação.

O ensino de CS1 possui importância por ser o primeiro contato do aluno com os princípios do pensamento sistemático e lógico, bem como a introdução à uma linguagem de programação, no entanto, um desafio recorrente são as altas taxas de evasão e

*Agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo processo 142476/2020-0.

reprovação [Bosse 2020, Walker 2017] que eles possuem. Dados ambos fatores, como propor soluções abrangentes para enfrentar essas dificuldades? Um possível primeiro passo é entender quais são de fato esses conteúdos abordados, de forma ampla.

Com motivação baseada nessas disciplinas introdutórias, bem como a potencial orientação que uma listagem abrangente dos tópicos mais abordados pode ter na criação de futuras intervenções em seu ensino e aprendizagem, a pergunta de pesquisa deste trabalho é: **PP**: *Quais os tópicos mais abordados em disciplinas de introdução à programação no ensino superior brasileiro?* De forma a responder esta PP, foi proposta uma consulta direta a meios públicos disponibilizados pelas IES, como os projetos pedagógicos de curso e ementários.

2. Trabalhos Relacionados

Os trabalhos considerados correlatos desta pesquisa são os que consolidam discussões a respeito dos tópicos ensinados em disciplinas de CS1, em especial, aqueles que apresentam alguma listagem de tais assuntos tiveram um foco maior.

Hertz e Ford [Hertz and Ford 2013] investigaram fatores que podem influenciar no ensino e aprendizagem de alunos em disciplinas de introdução à programação, envolvendo ambas CS1 e CS2. A lista de assuntos abordados foi derivada de um modelo de currículo para o ensino de computação em Artes Liberais (LACS, em inglês) [Kelemen et al. 2007], culminando em 17 tópicos. Os autores realizaram uma pesquisa online com docentes, verificando que as habilidades dos alunos estão quase sempre correlacionadas com a importância que o instrutor acredita que cada tópico ensinado possui.

Já Schulte e Bennedsen [Schulte and Bennedsen 2006] tentaram criar uma visão global de opiniões do que se deve ensinar em disciplinas de introdução à programação, objetivando uma reestruturação de CS1. A lista de assuntos abordados, com 28 tópicos, foi construída a partir dos resultados de outras pesquisas de opinião. A análise realizada pelos autores foi composta de fatores como a importância dos tópicos pelos docentes; o paradigma de programação ensinado; os tópicos que os alunos possuem maiores dificuldades; entre outros.

Por sua vez, Becker e Fitzpatrick [Becker and Fitzpatrick 2019] analisaram as ementas de disciplinas de CS1 de 916 instituições presentes no *QS World University Rankings* de 2016-2017[1]. O principal objetivo dos autores era responder o que exatamente os professores esperam de cursos introdutórios de programação, evidenciando 15 tópicos mais abordados, nomeados como resultados de aprendizagem.

3. Metodologia

A forma de identificar os tópicos mais abordados em disciplinas de CS1 nesta pesquisa foi idealizada em consulta direta em meios públicos disponibilizados pelas IES. Como esse processo de coleta precisaria ser manual e, considerando o vasto número de instituições, optou-se por analisar, em primeira instância, somente as Universidades Federais, utilizando uma lista base [Wikipédia 2021]. Além disso, os cursos de graduação em computação verificados foram os presentes nos Referenciais de Formação para os Cursos

[1] https://www.topuniversities.com/university-rankings/world-university-rankings/2016

de Graduação em Computação[2], organizado pela Sociedade Brasileira de Computação em 2017 [Zorzo et al. 2017], com exceção dos cursos superiores em tecnologia.

Para que uma disciplina de CS1 das IES fosse elegível, ela precisava atender dois critérios: possuir foco no ensino de conceitos de programação de computadores, considerando ensino de algoritmos em pseudocódigo e/ou linguagem de programação; e ser a primeira dessa categoria listada no sequenciamento de períodos sugeridos pela IES, ou seja, nenhuma disciplina que abordava os tópicos descritos no primeiro critério deveria ser cursada anteriormente. As informações escolhidas para identificação dos tópicos abordados em cada disciplina válida foram compostas pela *ementa* e *conteúdo programático*, embora esse segundo item não fosse obrigatório. A busca foi realizada através de documentos representando currículos oficiais e mais recentes possíveis em vigência da IES, como projeto pedagógico de curso, grade curricular, ementários, planos de aula ou página eletrônica da instituição.

A análise também foi realizada de forma manual: conforme novas disciplinas foram analisadas, novos tópicos foram identificados e adicionados à listagem resultante. Paralelamente, uma marcação foi assinalada correspondendo se uma disciplina abordava o tópico da lista resultante ou não, de modo a poder contabilizar todas as ocorrências quando o conjunto fosse completamente analisado.

4. Resultados e Discussão

Ao final da análise, que envolveu eventuais descartes de IES e disciplinas por não possuírem os requisitos, um total de 150 disciplinas de CS1 foram analisadas de 61 Universidades Federais, resultando em 63 tópicos encontrados, no entanto, como esta pesquisa buscou os tópicos mais abordados, um ranqueamento foi realizado com base no número de disciplinas em comum que esses tópicos eram mencionados nas ementas analisadas. Dessa forma, um primeiro corte foi feito, arbitrariamente, agrupando os tópicos que apareciam em pelo menos 10 disciplinas (6.7% do total), mas, para responder melhor a PP, um novo corte foi idealizado, com tópicos presentes em pelo menos um terço (50).

A Figura 1 mostra um resumo dos resultados dos processos de coleta e análise, já o ranqueamento dos tópicos está na Tabela 1, em que também foi realizada uma comparação da presença dos tópicos na disciplina de Algoritmos e Programação de Computadores (MC102)[3] da UNICAMP e nas listas obtidas nos trabalhos relacionados [Hertz and Ford 2013, Schulte and Bennedsen 2006, Becker and Fitzpatrick 2019]. Um exemplo de leitura desta Tabela seria que o tópico abordando *Comandos condicionais* apareceu em 138 disciplinas, possuindo cobertura de 92.0% do total de disciplinas analisadas. O mesmo tópico também é descrito em MC102 e nos trabalhos de Hertz e Ford (H&F), Schulte e Bennedsen (S&B), e Becker e Fitzpatrick (B&F).

Um fator importante a ser evidenciado é que embora esta pesquisa tenha se baseado nas ementas e conteúdos programáticos encontrados para obter os tópicos mais abordados nas disciplinas de CS1, não foi considerado que a ausência de algum deles

[2]Bacharelados em Ciência da Computação, Engenharia da Computação, Engenharia de Software e Sistemas de Informação, bem como Licenciatura em Computação/Informática.

[3]A ementa utilizada foi obtida em: `https://www.ic.unicamp.br/historico-ic/graduacao/programa-disciplinas/MC102-Algoritmos_e_Programacao_de_Computadores-INSTITUTO_DE_COMPUTACAO.pdf`

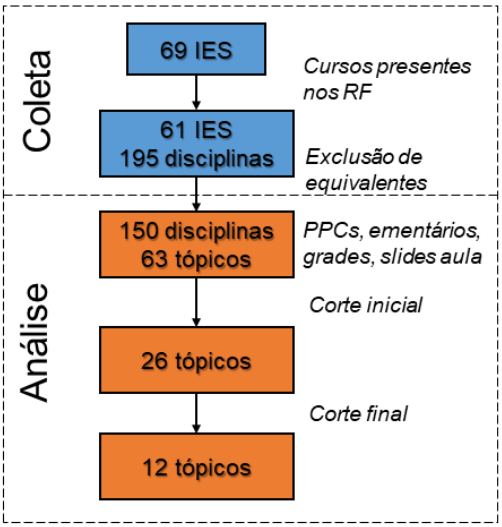


Figura 1. Resultados obtidos através dos processos de coleta e análise dos dados.

implique que eles não sejam ensinados nessas disciplinas, apenas indica que não foram encontrados nos meios consultados. Foi observado também que alguns listaram os tópicos de maneira muito simples, às vezes até repetindo alguns itens, em uma ordem que não necessariamente indicava a que é abordada em sala de aula; já outros deixam explícita a ordem a ser seguida, até mesmo com as horas-aula dedicadas a cada assunto. Além disso, é possível que alguns desses documentos possam não dar destaque, ou mesmo omitir, certos tópicos que consideram talvez triviais porque são necessários para os mais complexos que são descritos, sendo esta a possível causa que nenhum tópico esteja presente nas 150 disciplinas consultadas (Tabela 1), ou as demais discrepâncias observadas entre tópicos relacionados, como funções e passagem de parâmetros.

Comparando o ranqueamento dos tópicos mais abordados com a disciplina de MC102 e os trabalhos relacionados, também presente na Tabela 1, é possível perceber que todos os assuntos agrupados no corte final estão em MC102 e em pelo menos 2 dos 3 trabalhos relacionados utilizados. Esse resultado é importante, pois mostra correlação com as desenvolvidas em pesquisas alheias, sendo possível caracterizar como forma de validação do corte final desta pesquisa. Contudo, foi observado que alguns tópicos listados explicitamente separados aqui aparecem agrupados nas demais listas, por exemplo, *Comandos condicionais* e *Comandos de repetição* aparecem como *seleção e iteração* em [Schulte and Bennedsen 2006] e *construtos de controle* em [Hertz and Ford 2013]. A construção de todas as listas, inclusive a deste trabalho, levou em consideração a experiência dos pesquisadores, portanto, fatores de opinião podem estar presentes. Em sua maior parte essas ocorrências não foram consideradas como inconvenientes porque foi possível identificar a semelhança entre os tópicos, no entanto, algumas representações vagas também foram encontradas, como *Estruturas de dados simples (arrays, strings, ...)*, listados em [Schulte and Bennedsen 2006] e *Arrays, listas, vetores, etc.*, listados em [Becker and Fitzpatrick 2019]: a presença de indicadores que esses itens possuem outros não listados levanta a questão de quais tópicos seriam esses.

Tabela 1. Ranqueamento de tópicos mais abordados nas 150 disciplinas de CS1 analisadas. A ordenação foi realizada de forma decrescente pelo número de disciplinas (N) que os assuntos são abordados. A linha horizontal localizada aproximadamente na metade apresenta o corte final, evidenciando os tópicos que apareciam em pelo menos 50 disciplinas (33.3% do total).

Descrição do conteúdo abordado	Disc.	(%)	MC102	H&F	S&B	B&F
Variáveis, constantes e atribuições	139	92.7	✓	✓	✓	✓
Comandos condicionais	138	92.0	✓	✓	✓	✓
Comandos de repetição	135	90.0	✓	✓	✓	✓
Funções, modularização e subprogramas	129	86.0	✓	✓	✓	✓
Expressões aritméticas, lógicas e relacionais	128	85.3	✓	✓		✓
Variáveis compostas homogêneas unidimensionais	125	83.3	✓	✓	✓	✓
Variáveis compostas homogêneas multidimensionais	121	80.7	✓		✓	✓
Representações de algoritmos	114	76.0	✓		✓	✓
Entrada e saída de dados	95	63.3	✓	✓		✓
Variáveis compostas heterogêneas	90	60.0	✓		✓	✓
Recursão	62	41.3	✓	✓	✓	✓
Escopo de variáveis, passagem de parâmetros	59	39.3	✓	✓	✓	
Manipulação de arquivos	46	30.7	✓	✓		✓
Conceitos básicos de computadores	44	29.3	✓			
Ponteiros e alocação dinâmica de memória	44	29.3	✓		✓	
Algoritmos de busca	25	16.7	✓			
Algoritmos de ordenação	25	16.7	✓	✓		
Testes de código	22	14.7	✓	✓		✓
Depuração de código	18	12.0	✓	✓	✓	✓
Documentação de código	18	12.0	✓			✓
Ambiente de desenvolvimento	12	8.0			✓	
Aplicações da vida real	12	8.0				
Padrões de solução	12	8.0			✓	
Uso de bibliotecas e módulos	10	6.7			✓	
Histórico das linguagens de programação	10	6.7				
Programação orientada a objetos	10	6.7		✓	✓	✓

Alguns exemplos de tópicos presentes nos trabalhos relacionados, que não foram listados nesta pesquisa, envolvem conceitos mais abstratos como modelos mentais do computador, resolução de problemas e escrita de programas. Outros envolvem mais a orientação a objeto como classes e objetos, herança e polimorfismo, encapsulamento e design controlado por responsabilidade. Estruturas de dados avançadas (grafos, árvores, listas encadeadas) e eficiência de algoritmos também aparecem como tópicos dos trabalhos relacionados que não foram identificados nesta pesquisa ou não obtiveram frequência significativa para serem listados nos cortes estabelecidos na Tabela 1.

5. Conclusões e Próximos Passos

Os 12 tópicos representados pelo corte final desta pesquisa, que apareceram em pelo menos um terço do total de disciplinas analisadas, são compostos por conceitos de programação estruturada. Embora outros listados no corte inicial também façam parte deste paradigma, é possível que sua baixa frequência no conjunto analisado possa também estar relacionada com o fator de que somente a primeira disciplina de introdução à programação de cada IES foi verificada, ou seja, esses tópicos podem ser ensinados nas disciplinas seguintes.

É esperado que a listagem obtida nesta pesquisa possa ajudar, de forma abrangente, a justificar futuras intervenções no ensino e aprendizagem de CS1, no entanto, como apenas foram verificadas as Universidades Federais brasileiras, há espaço para expansão da análise incluindo demais instituições. Além disso, também é possível abordar outros fatores, como paradigma e linguagem de programação utilizada nos cursos de CS1.

Por fim, demais comparações desta lista com outras focadas em assuntos específicos, como tópicos com maior dificuldade também compõem trabalhos futuros. Em especial, os próximos passos desta tese envolvem o cruzamento com tópicos que surgem a partir de outras visões e intervenções do ensino e aprendizagem de CS1, como, por exemplo, a verificação de assuntos que possuem dificuldades que permanecem mesmo com o uso de ferramentas de correção automática de código.

Referências

Austing, R. H., Barnes, B. H., Bonnette, D. T., Engel, G. L., and Stokes, G. (1979). Curriculum '78: Recommendations for the undergraduate program in computer science— a report of the acm curriculum committee on computer science. *Commun. ACM*, 22(3):147–166.

Becker, B. A. and Fitzpatrick, T. (2019). What do cs1 syllabi reveal about our expectations of introductory programming students? In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, pages 1011–1017.

Bosse, Y. (2020). *Padrões de Dificuldades Relacionadas com o Aprendizado de Programação*. PhD thesis, Universidade de São Paulo.

Hertz, M. and Ford, S. M. (2013). Investigating factors of student learning in introductory courses. In *Proceeding of the 44th ACM Technical Symposium on Computer Science Education*, SIGCSE '13, page 195–200, New York, NY, USA. Association for Computing Machinery.

Kelemen, C. F., Consortium, L. A. C. S., et al. (2007). A 2007 model curriculum for a liberal arts degree in computer science. *Journal On Educational Resources In Computing*, 7(2).

Nascimento, P. B. d. (2018). Recomendação de ação pedagógica no ensino de introdução à programação por meio de raciocínio baseado em casos. Master's thesis, Programa de Pós-graduação em Informática. Instituto de Computação.

Schulte, C. and Bennedsen, J. (2006). What do teachers teach in introductory programming? In *Proceedings of the second international workshop on Computing education research*, pages 17–28.

Walker, H. M. (2017). Acm retention committee retention of students in introductory computing courses: Curricular issues and approaches. *ACM Inroads*, 8(4):14–16.

Wikipédia (2021). Lista de universidades federais do brasil — wikipédia, a enciclopédia livre. [Online; accessed 6-julho-2021].

Zorzo, A. F., Nunes, D., Matos, E., Steinmacher, I., de Araujo, R. M., Correia, R., and Martins, S. (2017). Referenciais de formação para os cursos de graduação em computação.

Segmentação Semi-Automática de Estruturas Torácicas em Exames de Tomografia

Ilan F. da Silva[1], **Alexandre X. Falcão**[1]

[1]Laboratory of Image Data Science (LIDS), Instituto de Computação,
Universidade Estadual de Campinas (Unicamp)

ilanfsilva@gmail.com, afalcao@ic.unicamp.br

Abstract. *The ALTIS method, developed by the research group, provides a fast and accurate automatic segmentation tool of the lungs and trachea in a chest computed tomography scan. Moreover, the method is robust, since its methodology was developed considering anomalies that deform the lungs. However, the method fails in severe cases of pulmonary anatomy impairment or in the presence of structures attached to the pleura, such as tumors or consolidations. The present work aims at exploring image processing techniques and graph search algorithms to perform a semi-automatic correction of the segmentation, in order to improve the ALTIS method for clinical use.*

Resumo. *O método ALTIS, desenvolvido pelo grupo de pesquisa, provê uma ferramenta rápida e acurada de segmentação automática dos pulmões e da traqueia em um exame de tomografia computadorizada do tórax. É, além disso, robusto, pois sua metodologia foi desenvolvida considerando possíveis anomalias que deformam os pulmões. Entretanto, o método falha em casos com alto grau de comprometimento da forma pulmonar ou na presença de estruturas, tais como tumores e consolidações, conexas à pleura. O presente trabalho visa explorar técnicas de processamento de imagens e algoritmos de busca em grafos em uma ferramenta para a correção semi-automática de segmentações, de modo a aperfeiçoar o ALTIS para uso clínico.*

1. Introdução

Doenças respiratórias estão entre as principais causas de óbito no mundo. Nesse contexto, a tomografia computadorizada (*computed tomography* – CT) é uma das mais importantes ferramentas para o diagnóstico e o posterior prognóstico de doenças respiratórias, auxiliando o médico especialista na tomada de decisões. No entanto, devido ao alto volume de imagens geradas e possíveis equívocos que um radiologista possa cometer ao analisá-las, ferramentas que auxiliam a análise de uma imagem volumétrica de CT são importantes para promover economia de tempo e diagnósticos mais precisos.

O primeiro passo para a análise computadorizada de exames de CT do sistema respiratório é a segmentação dos pulmões e da traqueia-brônquios. A segmentação consiste em duas etapas principais: detecção, que identifica os objetos de interesse, como é o caso da traqueia-brônquios e pulmões; e delineamento, com o propósito de definir as extensões espaciais desses objetos. Essa segmentação tem como objetivo principal a redução do espaço de busca e consequentemente a redução do esforço computacional de

métodos que farão uma análise mais detalhada, visando a detecção de lesões, extração de medidas, e sugestões de diagnóstico de doenças.

Nesse cenário, o *Automatic Lungs and Trachea Image Segmentation* (ALTIS) [Sousa et al. 2019] surge com a proposta de ser um rápido método capaz de segmentar e separar a traqueia e os pulmões direito e esquerdo. Dentre outros métodos que realizam a mesma segmentação e separação, o ALTIS é estatisticamente mais preciso e consideravelmente mais rápido em comparação com os métodos baseados em crescimento de regiões (PTK [Doel et al. 2012]), e os baseados em modelos probabilísticos de forma (SOSM-S [Phellan et al. 2016] e MALF [Aljabar et al. 2009]). Além disso, o ALTIS é inteiramente baseado em operadores de processamento de imagem que consideram a conectividade entre os elementos da imagem, livrando-o da necessidade de pré-treinamento em grandes bases de dados devidamente anotadas, escassas na área médica, como geralmente ocorre com os métodos baseados em Redes Neurais Convolucionais (CNNs).

Entretanto, em situações de alto grau de comprometimento dos pulmões ou na presença de estruturas, tais como tumores e consolidações, conexas à pleura, o método ALTIS pode resultar uma segmentação falha, como pode ser observado na Figura 1(a). Falhas nessas condições não são limitadas ao algoritmo do ALTIS, elas costumam ocorrer com os demais métodos. Essas falhas ocorrem devido à similaridade de intensidade entre as anomalias que comprometem os pulmões e tecidos adjacentes do corpo do paciente e pelo alto contraste entre tais anomalias e o parênquima pulmonar normal. Para contornar esse problema, este estudo propõe explorar algoritmos de busca em grafos para correção interativa de possíveis erros de segmentação. A proposta prevê ainda uma ferramenta computacional para a execução do ALTIS, visualização dos resultados, e correção interativa, quando necessária.

Para isso, entende-se o algoritmo do ALTIS sendo baseado em várias Transformadas Imagem-Floresta (IFT - *Image Foresting Transform* – uma metodologia para projeto de operadores de imagem por busca em grafos) [Falcão et al. 2004], visando estimar marcadores nos pulmões e traqueia-brônquios, bem como para delinear esses objetos. Desta forma, a segmentação final do ALTIS define cada objeto como uma floresta de caminhos ótimos enraizada em seus marcadores internos, e a versão diferencial do algoritmo da IFT (DIFT) [Falcão and Bergo 2004] pode ser usada com as mesmas características de imagem para a correção interativa de erros de segmentação. No algoritmo da DIFT, o usuário pode adicionar ou remover os marcadores da floresta anterior, gerando uma nova floresta onde os objetos resultam da união de árvores de caminhos ótimos enraizadas nos marcadores internos.

Além de viabilizar a correção, a ferramenta interativa pode ser executada a partir de marcadores selecionados pelo usuário, possibilitando a segmentação de outras estruturas do tórax não contempladas pelo ALTIS, assim como anomalias (consolidações e lesões) e sistema vascular. Para esse fim, poderá ser utilizado uma nova metodologia de IFT, Dynamic IFT [Bragantini et al. 2018], que calcula a floresta de caminhos ótimos com base nas informações dos caminhos, como textura, cor e formato, resultando experimentalmente em uma maior captura de pequenos detalhes como, por exemplo, os vasos sanguíneos do sistema vascular.

2. Metodologia

2.1. IFT

A Transformada Imagem-Floresta (*Image Foresting Transform* – IFT) é um arcabouço para o desenvolvimento de operadores de processamento de imagem baseados em conectividade ótima [Falcão et al. 2004].

Uma imagem pode ser definida como um par $\hat{I} = (D_3, \vec{I})$, onde $D_3 \subset \mathbb{N}^3$ é o domínio da imagem e $\vec{I}$ é a função que atribui um valor de intensidade para cada voxel $p \in D_3$. Essa imagem $\hat{I}$ pode ser interpretada como um grafo $G = (D_3, A)$, cujos vértices são os voxels e os arcos são definidos por uma relação de adjacência $A(p)$ dada por $A_\gamma : \{(p, q) \in D_3 \times D_3, \|q - p\| \leq \gamma\}$. Para um dado grafo G, a função critério f deve ser definida para qualquer caminho $\pi_q = \langle p_1, p_2, \ldots, p_n = q\rangle$, $(p_i, p_{i+1}) \in A$, $i = 1, 2, \ldots, n-1$, no conjunto Π_q de todos os possíveis caminhos com término q, incluindo os caminhos triviais $\pi_q = \langle q\rangle$. O algoritmo da IFT essencialmente minimiza um mapa de custos $C(q) = \min_{\forall \pi_q \in \Pi_q}\{f(\pi_q)\}$ por meio da partição do grafo em uma floresta de caminhos-ótimos P.

A IFT requer um conjunto de sementes $\mathcal{S} \subseteq D_3$ que servem como raízes para os caminhos ótimos. O algoritmo promove, então, uma competição ótima entre essas sementes de modo que cada semente conquista os voxels $p \in D_3 \setminus \mathcal{S}$ mais conectados de acordo com a minimização da função $f(\pi)$.

2.1.1. DIFT

A Transformada Imagem-Floresta Diferencial (*Differential Image Foresting Transform* – DIFT) [Falcão and Bergo 2004] é uma extensão do arcabouço original da IFT que permite a execução sequencial de IFTs de forma diferencial. Isto é, a cada nova iteração, sementes podem ser removidas e adicionadas para uma nova competição com as sementes remanescentes. Com isso, o algoritmo não precisa ser reexecutado completamente a cada momento. O custo computacional da DIFT é proporcional ao número de voxels que precisam ser atualizados, logo, o custo diminui conforme a correção progride.

O método ALTIS explora a IFT em todas as suas etapas, desde o cálculo dos marcadores internos em cada objeto de interesse (pulmão direito, pulmão esquerdo e traqueia) e externos indicando o fundo, até o delineamento propriamente dito. Portanto, a adaptação da segmentação gerada pelo ALTIS para a correção com a DIFT se torna viável e direta.

2.1.2. Dynamic IFT

Uma das formas de se definir a função critério de conectividade f do algoritmo da IFT é usar árvores dinâmicas que exploram a cor, textura e forma do caminho ótimo [Bragantini et al. 2018]. Para esse fim, o algoritmo do Dynamic IFT estende o algoritmo da IFT, mantendo árvores dinâmicas $\mathcal{T}_r$ para cada semente r pertencente ao conjunto de

sementes $\mathcal{S}$. Assim, define-se a função de conectividade f como

$$f_{max}(\langle q \rangle) = \begin{cases} 0, & \text{se } q \in \mathcal{S} \\ +\infty, & \text{caso contrário} \end{cases} \qquad (1)$$
$$f_{max}(\pi_p \cdot \langle p, q \rangle) = \max\{f_{max}(\pi_p), w(p, q)\}$$

em que w é definido como sendo o peso do arco associado aos vértices p e q, utilizado nesse trabalho como $w(p, q) = \|\mu_{\mathbf{R(p)}} - \mathbf{I(q)}\|$, com $R(p)$ sendo a raiz do caminho de p e $\mu_{\mathbf{r}}$ é equivalente ao vetor no espaço de cor das médias de todos os nós enraizados em r.

No entanto, para viabilidade de uso clínico, o Dynamic IFT deve possuir um algoritmo diferencial de modo a aproveitar, em uma sequência de interações, as florestas previamente calculadas, pois isso reduz drasticamente o tempo empregado entre as interações. Desse modo, será desenvolvido esse algoritmo e disponibilizado como uma nova rotina de segmentação na interface.

2.2. Correção por fechamento morfológico

Uma forma de incluir as anomalias conexas à pleura do pulmão é por meio de um simples fechamento morfológico. Essa abordagem é efetiva, mas em contrapartida há a inclusão do mediastino na máscara resultante de cada um dos pulmões. Para contornar esse problema, é realizado um fechamento morfológico exclusivo para a pleura não-mediastinal. Para tal, identifica-se os voxels da pleura mediastinal por uma dilatação geodésica na superfície do pulmão a partir dos voxels da intersecção entre a pleura e a traqueia-brônquios. A partir dessa identificação, elimina-se os voxels externos do fechamento morfológico que estão mais próximos dessa pleura mediastinal do que qualquer outro voxel pleural. Assim, o fechamento morfológico agrega as anomalias pleurais enquanto o mediastino permanece intacto.

3. Resultados Parciais

A metodologia apresentada neste presente trabalho foi implementada totalmente, exceto o desenvolvimento do algoritmo diferencial do Dynamic IFT.

Nesse sentido, a interface de visualização e segmentação disponibiliza a execução do método ALTIS para a imagem de CT que o usuário está lidando. Caso haja alguma falha de segmentação, é possível carregar e visualizar as sementes encontradas pelo método, como mostra a Figura 1(a), um pulmão com falha de segmentação devido ao padrão de vidro fosco. Logo em seguida, o usuário pode adicionar novas sementes por marcadores para o respectivo pulmão na região com anomalia, possibilitando uma rápida execução de uma IFT diferencial para corrigir e incluir a região na máscara de segmentação, ilustrado na Figura 1(b).

Opcionalmente, o usuário é capaz de realizar uma dilatação morfológica para corrigir automaticamente essa segmentação, mostrado na Figura 1(d). Nessa situação, não há necessidade de inserir marcadores nessas regiões errôneas. Dessa maneira, essa correção automática é conveniente em situações com diversas anomalias pulmonares, em que o processo de correção interativo pode ser longo e exaustivo ao usuário à medida que ele deve adicionar marcadores para cada região anômala.

Embora o algoritmo diferencial do Dynamic IFT não tenha sido desenvolvido, implementou-se o seu algoritmo original na interface de segmentação. Como pressupunha, o método de segmentação foi capaz de captar melhor pequenos detalhes da imagem original, como ilustra-se na Figura 2, em que o método segmentou com maior profundidade a traqueia, capturando os bronquíolos. Assim, é viável incorporar essa segmentação da traqueia-brônquios pelo Dynamic IFT com o resultado da segmentação dos pulmões pelo método ALTIS, promovendo uma segmentação mais precisa desses componentes do sistema respiratório.

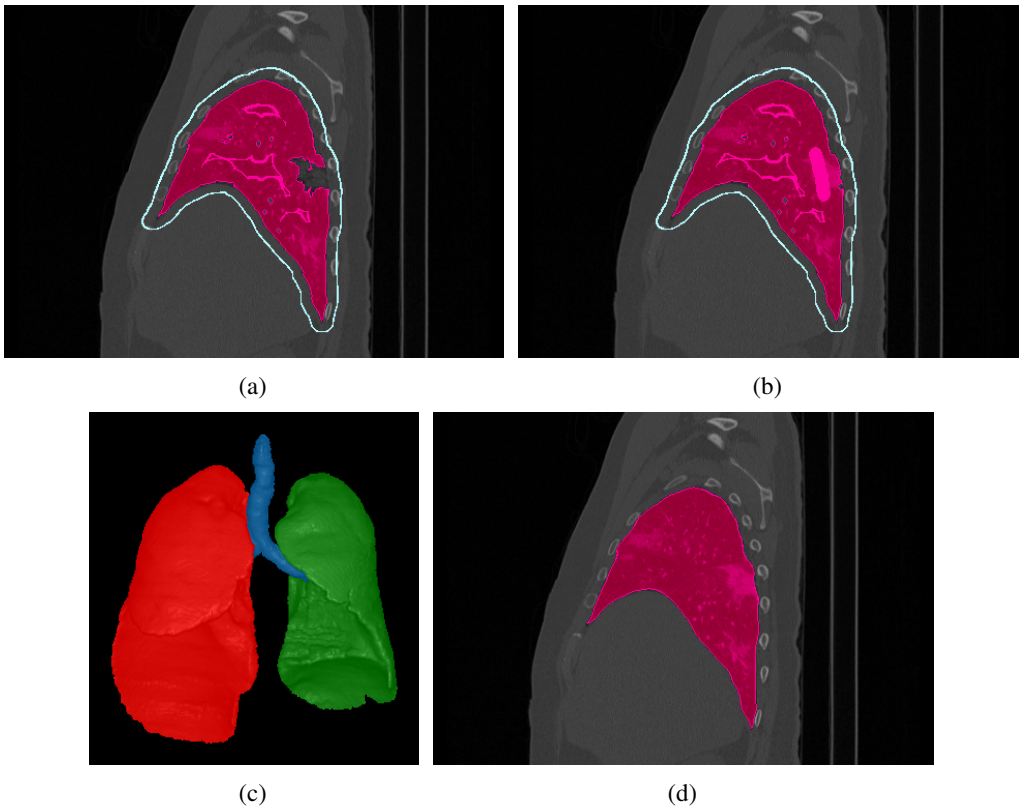


Figura 1. Fatia do plano sagital de uma imagem de CT de uma pessoa diagnosticada com Covid-19. (a) Resultado da segmentação pelo ALTIS em um dos pulmões, porém, devido ao padrão de vidro fosco houve uma falha de segmentação. Os traçados interno em vermelho e externo em ciano representam, respectivamente, as sementes internas e externas obtidas pelo método ALTIS. (b) Um marcador de pulmão foi adicionado à região com anomalia e então foi realizada a correção interativa executando o DIFT por meio da ferramenta de segmentação e visualização. (c) Renderização da segmentação retificada em (b). (d) Correção alternativa por fechamento morfológica.

4. Considerações finais

Portanto, nesse presente trabalho apresenta-se uma nova ferramenta de segmentação, visualização e renderização de imagens de tomografia computadorizada. Inclui-se na ferramenta a rotina de segmentação do ALTIS e bem como a correção interativa pelo DIFT. No entanto, ainda é necessário realizar a implementação diferencial do Dynamic

(a) (b)

Figura 2. Resultados de diferentes segmentações renderizadas da traqueia-brônquios em uma imagem de CT. (a) Resultado obtido a partir do ALTIS utilizando a segmentação por *watershed*. (b) Extensão da traqueia capturando bronquíolos com a segmentação do ALTIS utilizando Dynamic IFT.

IFT para conclusão do trabalho. Ademais, após finalização da ferramenta, será realizado uma avaliação subjetiva com médicos usuários da ferramenta.

Referências

Aljabar, P., Heckemann, R. A., Hammers, A., Hajnal, J. V., and Rueckert, D. (2009). Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *Neuroimage*, 46(3):726–738.

Bragantini, J., Martins, S. B., Castelo-Fernandez, C., and Falcão, A. X. (2018). Graph-based image segmentation using dynamic trees. In *Iberoamerican Congress on Pattern Recognition*, pages 470–478. Springer.

Doel, T., Matin, T. N., Gleeson, F. V., Gavaghan, D. J., and Grau, V. (2012). Pulmonary lobe segmentation from ct images using fissureness, airways, vessels and multilevel b-splines. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1491–1494. IEEE.

Falcão, A. X. and Bergo, F. P. (2004). Interactive volume segmentation with differential image foresting transforms. *IEEE Transactions on Medical Imaging*, 23(9):1100–1108.

Falcão, A. X., Stolfi, J., and Lotufo, R. A. (2004). The image foresting transform: Theory, algorithms, and applications. 26(1):19–29.

Phellan, R., Falcão, A. X., and Udupa, J. K. (2016). Medical image segmentation via atlases and fuzzy object models: Improving efficacy through optimum object search and fewer models. *Medical physics*, 43(1):401–410.

Sousa, A. M., Martins, S. B., Falcao, A. X., Reis, F., Bagatin, E., and Irion, K. (2019). Altis: A fast and automatic lung and trachea ct-image segmentation method. *Medical physics*, 46(11):4970–4982.