Análise da evolução dos títulos de pesquisas em computação da UFRJ por meio de *Differencial Tag Clouds*

Pedro Boechat
Engenharia de Computação e Informação
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brasil
pedroboechat@poli.ufrj.br

Pedro Kuchpil
Engenharia de Computação e Informação
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brasil
pedrokuchpil@poli.ufrj.br

***Abstract*—Este trabalho irá aplicar *Differential Tag Clouds* [1] para a visualização dos títulos de publicações relacionadas à Computação realizadas por professores e pesquisadores da Universidade Federal do Rio de Janeiro ao longo dos últimos 20 anos. Desta forma, será possível estabelecer termos e temas que se mantiveram como protagonistas durante este período, e também identificar novas e velhas tendências na pesquisa dentro da área.**

I. Introdução

A Universidade Federal do Rio de Janeiro (UFRJ), ao longo de seus 101 anos, se consolidou como uma das melhores universidades da América Latina [2], em parte por conta da grande produção acadêmica de seus pesquisadores. Programas como o Programa de Engenharia de Sistemas e Computação (PESC), entre tantos outros da UFRJ, rotineiramente são classificado com a nota máxima da CAPES [3], provando a competência e a importância da universidade para a pesquisa brasileira.

Com o intuito de analisar a evolução da pesquisa na UFRJ ao longo dos anos, este trabalho se utilizará de dados do Currículo Lattes de pesquisadores da universidade para responder às seguintes perguntas:

1) Quais foram os tópicos mais pesquisados dentro das áreas de computação?
2) De que forma ocorreu a mudança desses tópicos ao longo dos anos?

II. Dados

Os dados utilizados foram extraídos do Currículo Lattes de professores da universidade. A base de dados possui 70.740 publicações, contendo DOI, título, áreas de conhecimento, idioma e ano da publicação, e foi disponibilizada pelo Conecta UFRJ [4].

Após análise inicial, foi constatado que pouco menos de 5% dos dados não possuia título ou ano da publicação e 18% não possuia nenhuma área de conhecimento. Assim, começamos a manipulação dos dados tratando essas inconsistências, além de removendo quebras de linha, *tags* HTML e entradas nulas dos títulos. Ao analisar o idioma dos artigos, encontramos que cerca de 37% não possuíam o idioma e mais de 60% eram escritos na língua inglesa.

Considerando as áreas de conhecimento que possuíam mais publicações, vimos que grande parte delas eram áreas dos campos de medicina, biologia e química.

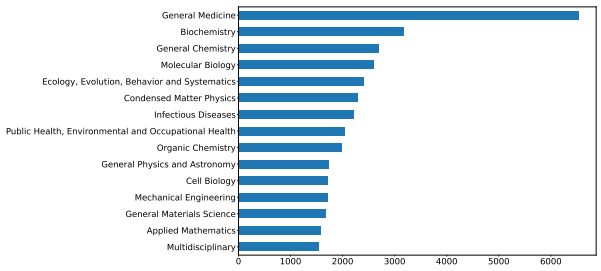

Fig. 1. 15 áreas de conhecimento com mais publicações.

A fim de esclarecer a mudança do comportamento dos títulos de publicações ao longo dos anos, este trabalho se debruçará apenas sobre registros de áreas do conhecimento ligadas à Computação, que são 2131 na base, para que a variação dos títulos de publicações de um período específico sejam melhor percebidas quando comparadas à coleção geral. Além disso, só serão analisados artigos escritos em língua inglesa, que são maioria, para que o pré-processamento também possa ser o mesmo para todos os dados. Logo, geramos o mesmo gráfico para essas áreas.

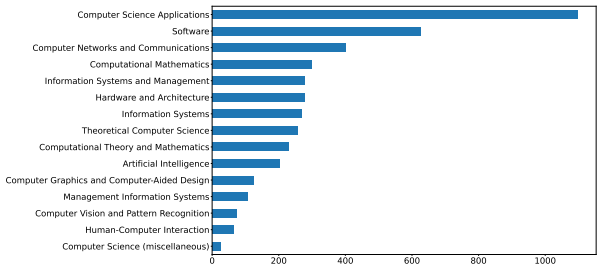

Fig. 2. Áreas de conhecimento relacionadas à computação.

Também analisamos o número de publicações por ano e foi possível notar que as publicações de áreas de conhecimento relacionadas à computação cresceram de forma similar ao número total de publicações.

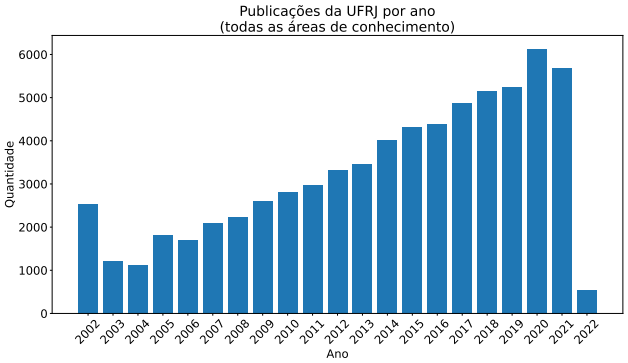

Fig. 3. Artigos publicados por ano em todas as áreas de conhecimento.

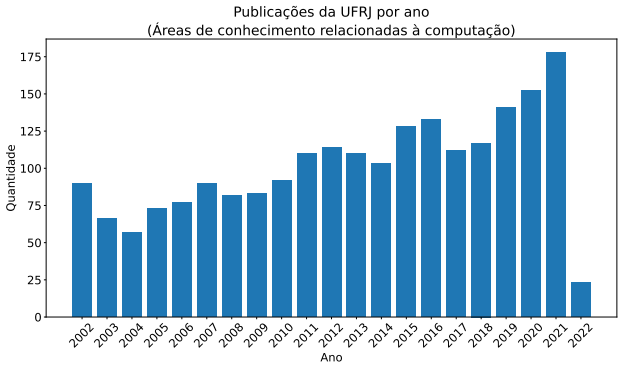

Fig. 4. Artigos publicados por ano em áreas de conhecimento relacionadas à computação.

III. Metodologia

O trabalho foi desenvolvido em *Python*, carregando os dados disponibilizados como *JSON* para um dataframe do Pandas [5].

Nos registros trabalhados, muitos dados do idioma da publicação estavam como nulos, ou então ainda erroneamente classificados. De modo a corrigir essa informação, foi utilizada a biblioteca `fasttext` [6], que, com um auxílio de um modelo pré-treinado, é capaz de fazer essa classificação. Apenas registros cujo resumo havia probabilidade maior do que 90% de ser em inglês, de acordo com o modelo, foram considerados na análise.

Após a remoção dos registros que não seguiam as regras estabelecidas, foi realizado o processamento do campo de título dos artigos, com a remoção de *stopwords*, para que fossem considerados apenas termos relevantes à publicação. O *Stemmer* de *Porter* [7] também foi aplicado, a fim de normalizar termos.

Continuando os passos descritos por [1], é calculado o *TF-IDF* dos termos, com base no ano de sua publicação. Para isso, foi criado uma coleção de documentos, um para cada intervalo de 3 anos (a partir de 2002) e contendo os títulos já processados de todas as publicações relevantes para o estudo datadas do respectivo intervalo.

Assim, é possível calcular a *Summary Tag Cloud*, que seleciona os 40 termos mais bem classificados na coleção de documentos. Ela está presente na Figura 5.

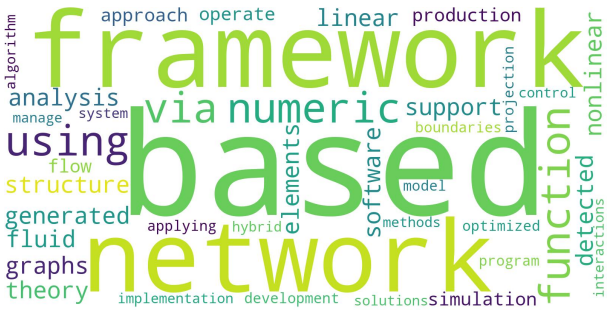

Fig. 5. *Summary Tag Cloud* da coleção de títulos de artigos

Após essa etapa, a *Differential Tag Cloud* pode ser calculada para os intervalo de anos, considerando as palavras mais bem colocadas que não estão na nuvem de *summary*. Desta forma, podemos identificar os temas e as tendências que marcaram um período específico. As nuvens geradas estão nas figuras de 6 a 11.

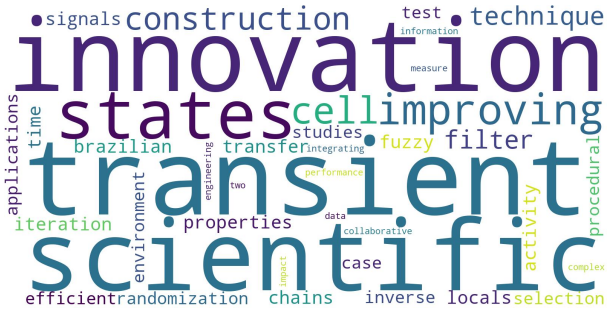

Fig. 6. *Differencial Tag Cloud* de 2002 a 2004

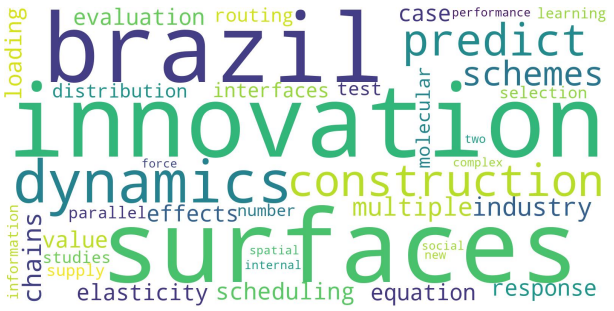

Fig. 7. *Differencial Tag Cloud* de 2005 a 2007

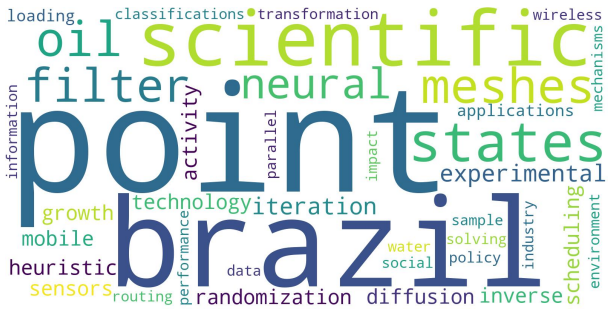

Fig. 8. *Differencial Tag Cloud* de 2008 a 2010

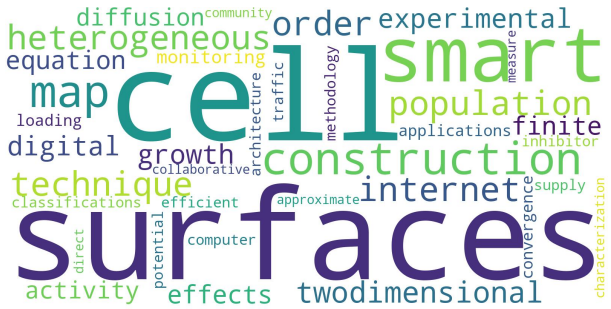

Fig. 9. *Differencial Tag Cloud* de 2011 a 2013

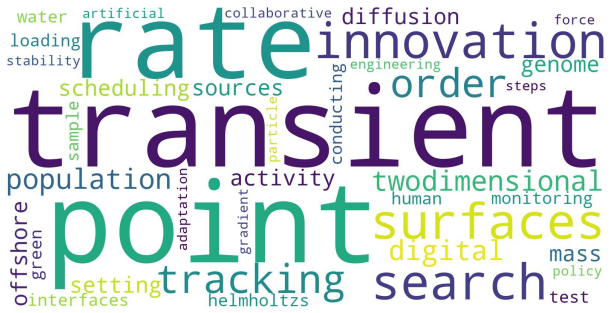

Fig. 10. *Differencial Tag Cloud* de 2014 a 2016

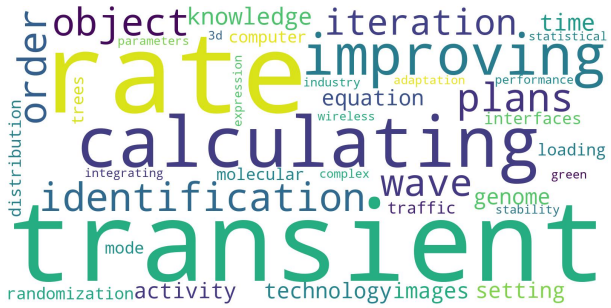

Fig. 11. *Differencial Tag Cloud* de 2017 a 2019

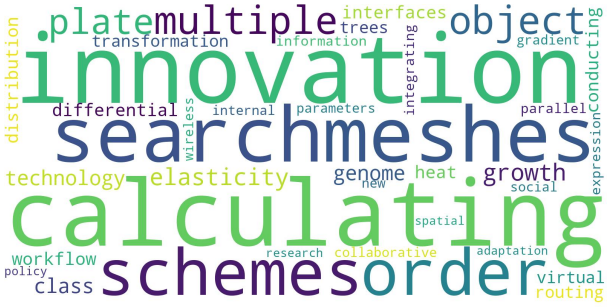

Fig. 12. *Differencial Tag Cloud* de 2020 a 2022

IV. Resultados e Conclusões

Como é possível observar nas imagens, ocorreram algumas mudanças significativas nos principais temas pesquisados na área de computação dentro da UFRJ ao longo do período analisado. Inicialmente, eram mais comuns publicações sobre aplicações científicas e sinais. Já entre 2008 e 2010, são presentes temas consideravelmente importantes, como pesquisas relacionadas a petróleo (pouco tempo depois da descoberta do Pré-Sal) e a redes neurais. Atualmente, nos três últimos anos, vale destacar os estudos sobre políticas, temas sociais e paralelismo. Cada um a sua maneira, são temas muito relevantes para a computação.

A criação de *Differencial Tag Clouds* para títulos de publicações da UFRJ permite acompanhar a evolução da pesquisa pública de excelência, sobre que temas já foram abordados e quais tendências estão sendo seguidas (ou criadas) dentro da instituição. Este trabalho sobre a área de computação pode ser replicado para qualquer outra, ou então ainda para a universidade como um todo.

Como trabalho futuro, cabe pensar em um gerador destas nuvens, analisando os dados públicos disponibilizados no Lattes dos pesquisadores de todo o Brasil. Cabe considerar também que diversos termos foram encontrados repetidas vezes, o que poderia ser evitado aumentando o tamanho da *Summary Tag Cloud*, ou então impedindo que palavras sejam adicionadas se estiverem em mais de um determinado número de nuvens diferenciais.

References

[1] Xexeo, G., Morgado, F., & Fiuza, P. (2009). Differential Tag Clouds: Highlighting Particular Features in Documents. 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology.
[2] https://cos.ufrj.br/
[3] https://www.timeshighereducation.com/world-university-rankings/federal-university-rio-de-janeiro
[4] https://www.conecta.parque.ufrj.br/
[5] https://pandas.pydata.org/
[6] https://fasttext.cc/
[7] https://tartarus.org/martin/PorterStemmer/