An analysis of the evolution of computation research titles at UFRJ using 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*—This work applies Differential Tag Clouds [1] for the visualization of publication titles of teachers and researchers from the Federal University of Rio de Janeiro in computing-related areas on the last 20 years. In this way, it will be possible to establish terms and themes that maintened themselves as protagonits during this period, and also identify old and new tendencies of research inside this area.**

I. Introduction

The Federal University of Rio de Janeiro (UFRJ, in its portuguese acronym), consolidated itself as one of the top universities in Latin America during its 101 years of history, in part due to the great academic production of its researchers. Programs like the Systems and Computing Engineering Program (PESC) and many others frequently are given the highest possible score by CAPES, the governmental agency responsible for classifying graduation courses, proving UFRJ's compentence and importance to brazilian research.

In order to analyze the evolution of the research conducted by UFRJ during the years, this project will utilize data available in the *Lattes* Curriculum of researches to answer the following questions:

1) Which topics were more present in publication titles considering computing-related subjects?
2) How the presence of these topics changed according to time?

II. Data

The data utilized was extracted from the *Lattes* Curriculum of researches of the university. The database has 70,040 publications, with DOI, title, subject, languange and publication year, and was disponibilized by *Conecta UFRJ* [4]

After an initial analysis, it was noted that less than 5% of the data did not have an title or publishment year and 18% did not have any subject. We started the data manipulation to treat those inconsistencies, and also removing new lines and HTML tags. We also found that 37% of the records did not have the subject, and more than 60% of our base was written in english.

Considering the subjects of the publications, we noted that great part of it belonged to the fields of medicine, biology and chemistry.

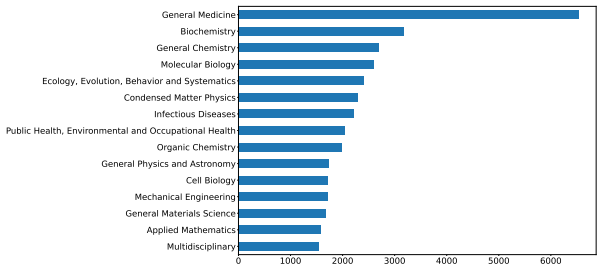

Fig. 1. 15 subjects with more publications.

To clarify the changes of the publication titles according to time, this project will only focus on records with subjects connected to computing, which are 2,131 in the database, so that the variation of the titles of a specific period is better perceived when compared to the general collection of titles. Also, only publications written in english will be analyzed, in order to keep the processing of the data the same for every record considered. We generated the same graphics to this specific area.

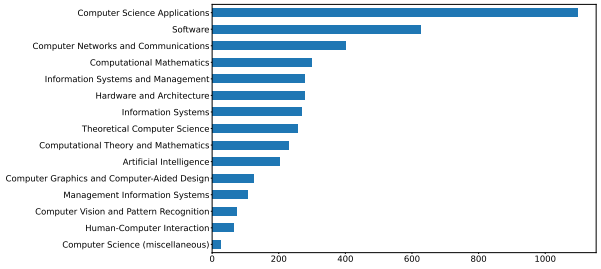

Fig. 2. Subjects related to computing.

We also analyzed the number of publications by year and it was possible to note that the ones with computing subjects had a similar growth when compared to the total number of publications.

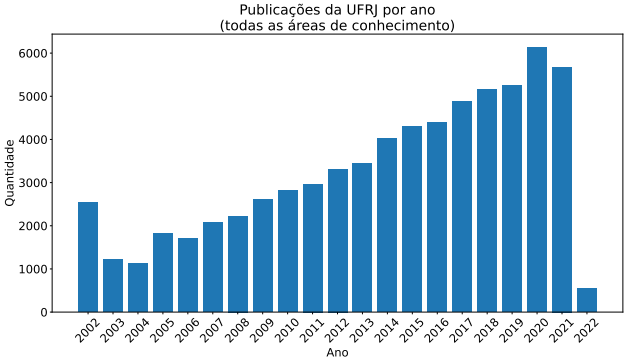

Fig. 3. Publications by year in every subject.

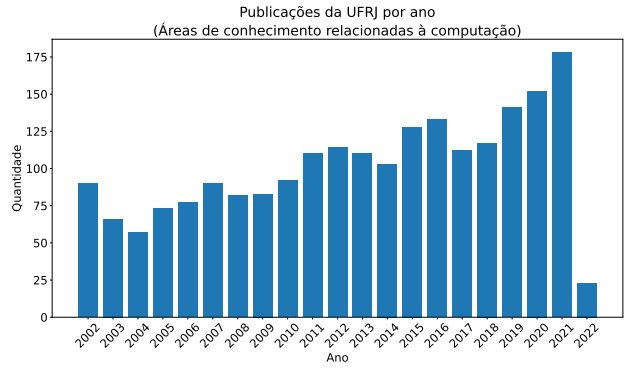

Fig. 4. Publication by year in computing-related subjects.

III. Metodology

The work was developed in *Python*, loading the *JSON* dataset as a Pandas [5] dataframe.

The publication language tag for many of the entries in the dataframe were missing or wrongly classified. To try improving this situation, it was used the `fasttext` library [6], which classifies the publication language by its abstract with a pre-trained model. Only classifications with confidence of over 90% of being in English were considered in this analysis.

After the removal of entries that were invalid through our stablished rules, we processed the title field, removing stopwords so it would only contain relevant terms for our analysis. We also applied the Porter Stemmer [7] to normalize the terms.

Continuing the steps described in [1], we calculated the *TF-IDF* of the terms, based on the publication year of the articles. For this step, it was created a collection of documents, one for each 3 year interval (beginning in 2002) and containing the processed titles for all the relevant publications for the study of each interval.

This way, it was possible to calculate the Summary Tag Cloud, which selects the 40 best classified terms in the document collection. It is present in Figure 5.

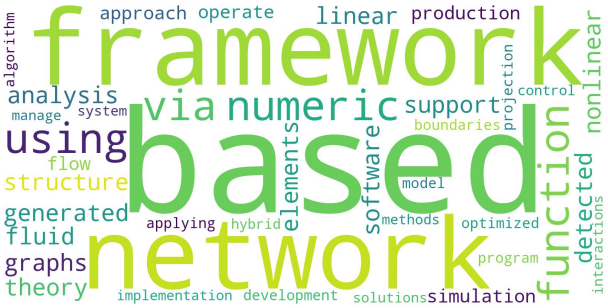

Fig. 5. *Summary Tag Cloud* for the whole collection of article titles

After this step, a Differential Tag Cloud can be calculated for each of the year intervals, considering the best ranked words that are not in the summary wordcloud. With those we could identify the subjects and trends that marked an specific period. The generated wordclouds are in the Figures 6 to 11.

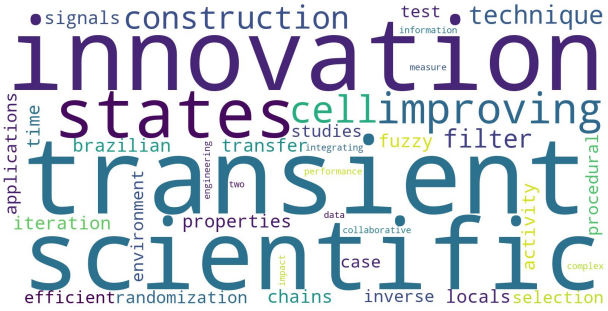

Fig. 6. *Differencial Tag Cloud* from 2002 to 2004

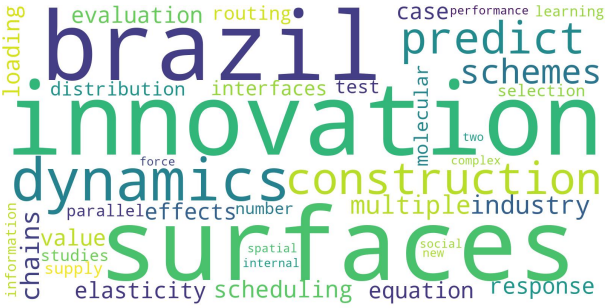

Fig. 7. *Differencial Tag Cloud* from 2005 to 2007

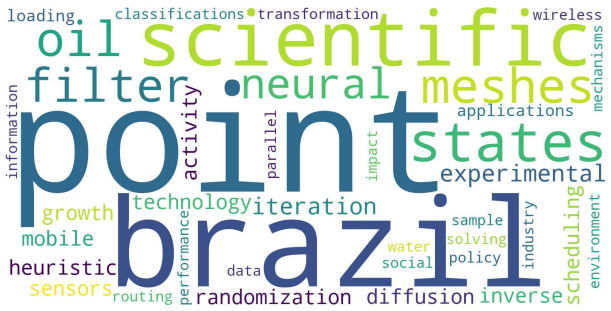

Fig. 8. *Differencial Tag Cloud* from 2008 to 2010

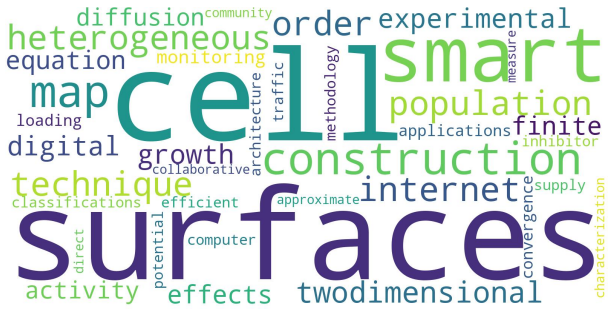

Fig. 9. *Differencial Tag Cloud* from 2011 to 2013

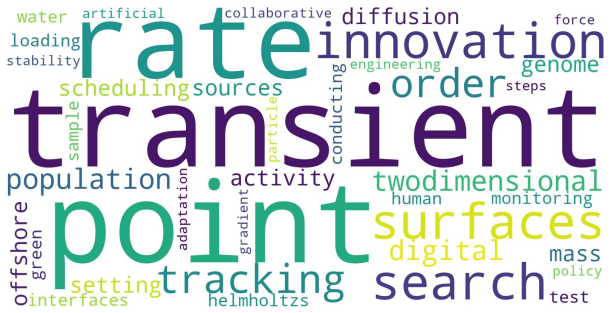

Fig. 10. *Differencial Tag Cloud* from 2014 to 2016

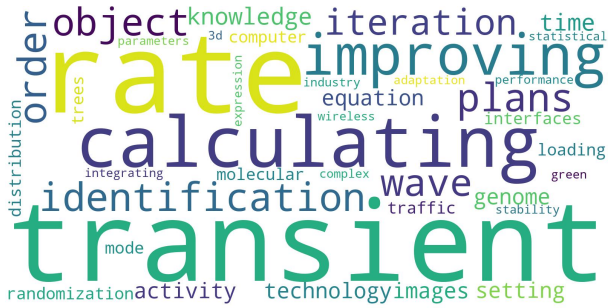

Fig. 11. *Differencial Tag Cloud* from 2017 to 2019

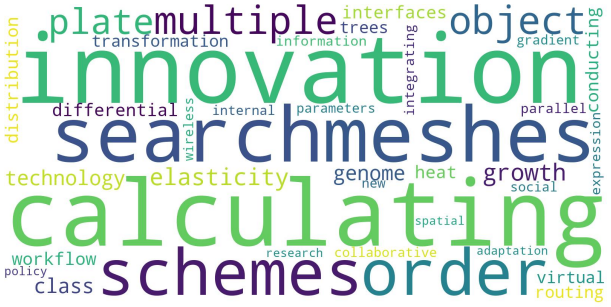

Fig. 12. *Differencial Tag Cloud* from 2020 to 2022

IV. Results and Conclusions

As it is possible to observe in the images, some significant changes happened in the main themes researched related to the computing area during the considered period. Initially, publications about scientific applications and signals were more common. During 2008 until 2010, important themes like oil (as the pre-salt layer of oil was discovered in Brazil around that time) and neural networks are observed. Currently, during the last three years, politics, social studies and parallelism can be highlighted. Each in its own way, all are very relevant to the discussions happening in computing research.

The creation of Different Tag Clouds to the titles of UFRJ publications allows the observation of the evolution of the public research of excellence happening in Brazil. With them, it is possible to see which themes already were the most common and which are the tendencies being followed (or created). In the Summary Tag, it is possible to point the main themes present in the general collection, the ones that never left the papers published by the researches. This project about the computing area can be replicated to any other, or even to the university as a whole.

As a future work, it is possible to think in a generator of this clouds, analyzing public data available in the *Lattes* Curriculums of researches of all Brazil. We can also consider that many terms were repeated across the Differential Cloud Tags, which could be avoided by enlarging the size of the Summary Tag Cloud, or preventing words to be added if they were to be present in an elevated number of differential clouds.

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/