PhD Thesis Report

Title of the PhD thesis: "**Measuring the Novelty of Scientific Papers**"

Author of the PhD thesis: **Pavel Savov**

Polish-Japanese Academy of Information Technology, 2021

Research Topic and Contribution of the Thesis

The thesis of Pavel Savov addresses the problem of the automatic processing of scholarly publications with the aim of measuring the innovativeness of papers. This topic is one of the core research problems in Scientometrics. The body of work presented in this thesis makes an important contribution to this field.

The proposed methods are novel. They rely on the full text content analysis of scientific corpora in a diachronic perspective. Pavel Savov defines the innovation score of papers as a real number metric that measures the novelty of a paper by taking into account its similarity with past and future papers.

Previous existing methods that try to characterize various aspects of the innovative quality of papers do not take into consideration the text content of papers but rely solely on the citation graphs. In such methods, factors such as the authors' or institutions' reputation, early citations, self-citations, can favour some particular papers, while other papers can express innovative ideas but get fewer citations. In this perspective, Pavel Savov's work brings an innovative approach that overcomes the shortcomings of the previous methods. The main argument of Pavel Savov's approach is the comparison between the publication year of a paper and its predicted publication year, where the predictions are made analysing the full text content of papers. The innovation score of a paper is obtained by comparing the sets of topics present in the paper with those of the past and future papers. From a technical point of view, the innovation score is calculated based on the output of state-of-the-art topic modelling approaches, firstly SVM classifiers with Latent Dirichlet Allocation (LDA), and secondly an ordinal regression model with Correlated Topic Models (CTM), and finally using BERT embeddings to predict the age of sentences and then the age of the entire document.

To validate this new approach, Pavel Savov conducts several experiments on corpora of scientific papers published over periods of more than 20 years, namely the World Wide Web Conference (WWW), the International ACM SIGIR Conference, and the Journal of Artificial Societies and Social Simulation (JASS). He shows how the measurement of the novelty of papers can be used to identify

breakthrough papers and years, and to predict document age. The comparison between the innovation score and citation counts shows that the innovation scores do no correlate with citation counts, although high citation counts tend to imply high innovation scores. Thus, the quality of innovativeness of a paper is somehow independent from the citation counts and the proposed methodology can be considered as complementary to other existing methods that rely on citations.

Thesis Content

The thesis is composed of an introduction, a literature review, three peer-reviewed papers, and a conclusion.

The introduction states the main research problem and its motivations. The notion of novelty is put in relation with the concept of the merit of publications, which is distinguished from impact. The main objective of this research, which is the identification of novelty, is defined within the field of Scientometrics. The author clearly identifies the biases that can exist in the identification of novelty in the existing approaches based on citation counts, and in traditional Scientometrics. He enumerates the major contributions of the thesis that consist in the definition of the paper innovation score, and two methods for dating scientific papers.

The Literature Review, chapter 2, gives a concise description of the existing metrics that are used in Scientometrics and Altmetrics. This chapter also outlines the most important methods used in the field of mining scientific papers: co-citation analysis, topic modelling, various text mining approaches, trend detection. The author pays specific attention to the works that are related to the identification of breakthroughs and sleeping beauties, and also document dating.

Chapter 3 presents the **paper "Identifying Breakthrough Scientific Papers"** published in *Information Processing & Management*, 2020. In this paper, the authors use Latent Dirichlet Allocation (LDA) topic model to identify topics in all papers and years. The number of topics has been calculated by maximizing the topic coherence measure. This paper introduces the innovation score of a paper as a real number measure, calculated as the weighted mean classification error with respect to the publication years. To compare papers published on different years, the innovation score is adjusted to account for the publication year of a paper with respect to the minimum and maximum publication year in the data set. The results show that there does not exist a strong correlation between the innovation scores of papers and citation counts. This chapter includes many examples and also provides the full list of topics that have been identified. The discussion addresses thoroughly some of the limitations of this study, especially the fact that the innovation score is less suitable for recent papers, because of the lack of future data.

Chapter 4 presents the paper **"Innovativeness Analysis of Scholarly Publications by Age Prediction using Ordinal Regression"**, published in the *International Conference on Computational Science*, 2020. This paper improves on the previous study by applying Correlated Topic Models (CTM) instead of LDA, and ordinal regression instead of SVM classifiers. The results show that these new methods improve the year prediction accuracy for both data sets of WWW and JASSS. After taking into

consideration the existing citation counts for these papers, this study confirms the previous result that innovation scores are only weakly correlated with citation counts.

Chapter 5 presents the paper **"Predicting the Age of Scientific Papers"**, published in the *International Conference on Computational Science*, 2021. This paper introduces a new method to predict sentence age, using state-of-the-art SciBERT embeddings. Then, document ages are predicted by the aggregation of sentence age predictions. This approach allows to improve the existing methods for document age prediction, and achieves a mean prediction error of about 0.6 years using the weighted mean with the sentence offset and the weighted mean with TextRank.

The Conclusion of the thesis provides a brief summary of the main contributions and outlines some of the future work.

Discussion on the Content and the Results

One major difficulty in tackling the problem of the identification of novelty in science is the fact that there does not exist an operational definition of novelty or innovativeness of papers or any other research outcomes in general. Pavel Savov provides such a definition, expressed by the innovation score of papers that expresses the extent to which a paper contains topics that tend to be absent from previous publications and present in future ones. This definition and its experimental validation constitute the main contribution of this thesis.

One important advantage of the proposed methodology is the fact that all of the processing steps are completely automated, which means that this approach can be easily applied to large collections of documents. However, as Pavel Savov points out, the innovation score of a paper depends to a large extent on the time span of the data set. In fact, the correct measurement of novelty requires hindsight: the innovation score is not suitable for recent papers.

The third paper in the thesis addresses the problem of document age prediction. The proposed methodology is complementary to that of the first two papers, namely the experiments related to the innovation score. However, these two types of studies are presented as independent from one another, and the thesis does not elaborate on the possible links between the two. Only one paragraph on the future work mentions this issue.

The solutions proposed in this thesis are technologically mature and rely on state-of-the-art models and approaches in Natural Language Processing (NLP). Overall, the work of Pavel Savov demonstrates an in-depth knowledge and understanding of the state of the art, both in NLP and Scientometrics. The proposed definitions and novel methods are original and the analysis is compelling. Furthermore, much attention is paid to the rigorous description of the experiments and the results, which contributes to the general quality of this work.

Conclusion

The thesis of Pavel Savov makes a substantial and sound contribution of the field of Scientometrics, and more precisely to the automatic processing of the full text of scholarly publications. This work provides automated methods to tackle the complex problem of novelty detection which is one of the fundamental objectives of Scientometrics. The approach is original, substantially documented and validated by the experimental results.

I find that this work qualifies as a PhD thesis and I strongly support its defence.

Besançon, March 4th, 2022

Dr Iana Atanassova

CRIT, Université de Bourgogne Franche-Comté

30 rue Mégevand 25000 Besançon, France

Institut Universitaire de France (IUF)