Trend analysis of scientometric indicators

Muhamedyev R I, Makhambetali T K

Satbayev University, Kazakhstan, Almaty, Satpayev str., 22A

Author's e-mail: ravil.muhamedyev@gmail.com, makhambetaliev96@gmail.com


Abstract

In this paper, we consider scientometric indicators of such a rapidly developing field of research as automatic text processing (natural language processing). Differential indicators of speed and acceleration were used to assess the dynamics of the development of NLP domains. The assessment was based on data from a direct bibliographic database Science. Calculations were performed for the following NLP subdomains: grammar checking, information extraction, text categorization, dialogue systems, speech recognition, machine translation, information search, answers to questions, opinion analysis, intelligent advisers and others. Areas with high growth rates (grammar checking, dialog systems, deep learning) and areas that lost the preexisting dynamics of publication activity growth (automatic summarization, speech recognition, information retrieval) were identified.The proposed indicators allow to visually express changes in the dynamics of scientometric indicators, which may beuseful in assessing the prospects of research areas.

Keywords: bibliometric, scientometrics, natural language processing, citations, differential indicators, D1, D2

1 Introduction

The field of research, combined terms of natural language processing (NLP) or automatic word processing, has caused great and constantly growing interest. Recently, scientific research and a general increased level of calculations have led to a number of breakthrough results in NLP, among which are achievements in the field of machine translation, automatic summation, information retrieval, answer to questions and mood analysis [1]. Selecting particular domainsit is possible to consider how the interest of researchers changes over time, and those areas of research that attract particular attention can be revealed. The number of publications in almost all areas of natural language processing is increasing.It is not enough to predict change of interest in particular domain using only number of publications. Differential indicators that were introduced in [2] are necessary to assess the speed and acceleration of changes in bibliometric indicators which help us to judge more precisely about the possible interest of particular domain in the future. In turn, speed and acceleration may indicate an increase or decrease in the interest of researchers in individual NLP subdomains. In this paper, the number of publications and the number of citations are consideredin differential indicators.

2 Methods and Data

The data was collected and attributed from the point of view of the tasks to be solved as "**NLP tasks**" or "**Tasks**" group as it was introduced in [3] and then distributed to the number of methods ("**Scientific NLP Methods**" or "**Methods**" group) such as: Machine Learning, Neural Networks, Deep Learning, Fuzzy Logic, First order logic, Knowledge representation, Evolutionary computation & Genetic programming, Rule based system, Unsupervised learning, Clustering, Supervised learning, Statistical methods, Bayesian networks, Semantic c networks, Keyword Spotting, Lexical affinity, Ontology, Information fusion, Taxonomy [2].To assess the dynamics of changes in publication activity indicators D1(speed) and D2(acceleration) [2] are used and defined as follows:

$$D1_i^j(t_k) = \beta \times \frac{dn_i^j(t_k)}{dt} + \gamma \times \frac{dc_i^j(t_k)}{dt},$$

$$D2_i^j(t_k) = \beta' \times \frac{d(dn_i^j(t_k)/dt)}{dt} + \gamma' \times \frac{d(dc_i^j(t_k)/dt)}{dt},$$

where n_i and c_i are number of publications and citations respectively, determined using search term $j, \beta, \gamma, \beta', \gamma$ some empirical coefficients that regulate the "weight" of the contribution of the number of publications, the speed and the acceleration of number of publications n_i and the speed and acceleration of number of citations c_i respectively.

In present work, D1 and D2 are calculated separately for publication and citation number. We assume β, γas equal to 1. Due to the peculiarities of the numerical calculation of derivatives, indicators D1 and D2 can only be calculated for previous years. However, we are interested in assessing the dynamics of changes in these indicators in the future, 1 or 2 years in advance. For this, using annual data on publication activity, regression models are constructed. As it is known know, the cost function of the regression model is described by an expression of the form:

$$J(\theta) = \min \quad \frac{1}{2m} [\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2],$$

where m – amount of data, and hypothesis function:

$$h_\theta = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \cdots + \theta_n x^n,$$

regression parameters - $\theta_i \in \theta$ and regularization parameter - λ. In this case m is a value of BI in the moment 1 to m. The minimization of the cost function is performed by one of the gradient descent algorithms: *Conjugate gradient*, *BFGS*, *L-BFGS*. Having the hypothesis function, we calculate additional values for the number of publications and citations both in the intervals between the available annual values and the predicted values.

3 Results

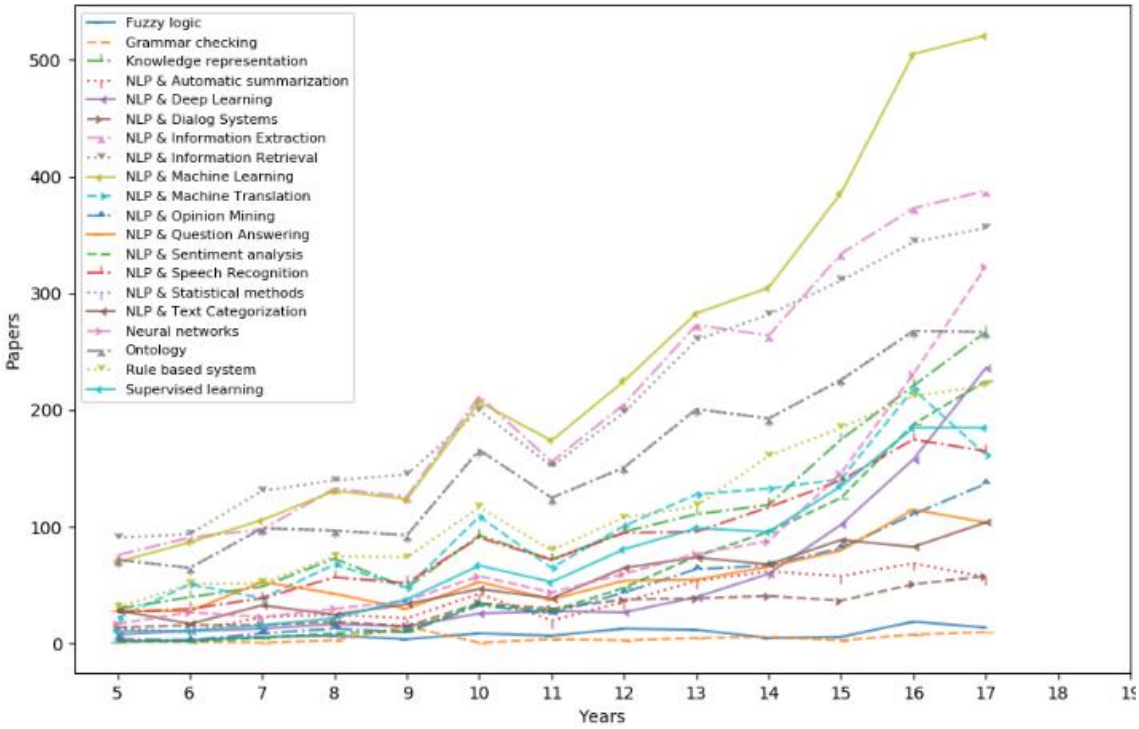

Figure 1 Publication activity

Based on the data above regression was built (Figure 2) in order to make curve more smooth and get additional predicted data for next couple years. Degree of regression was chosen in such way that minimizes mean squared error.

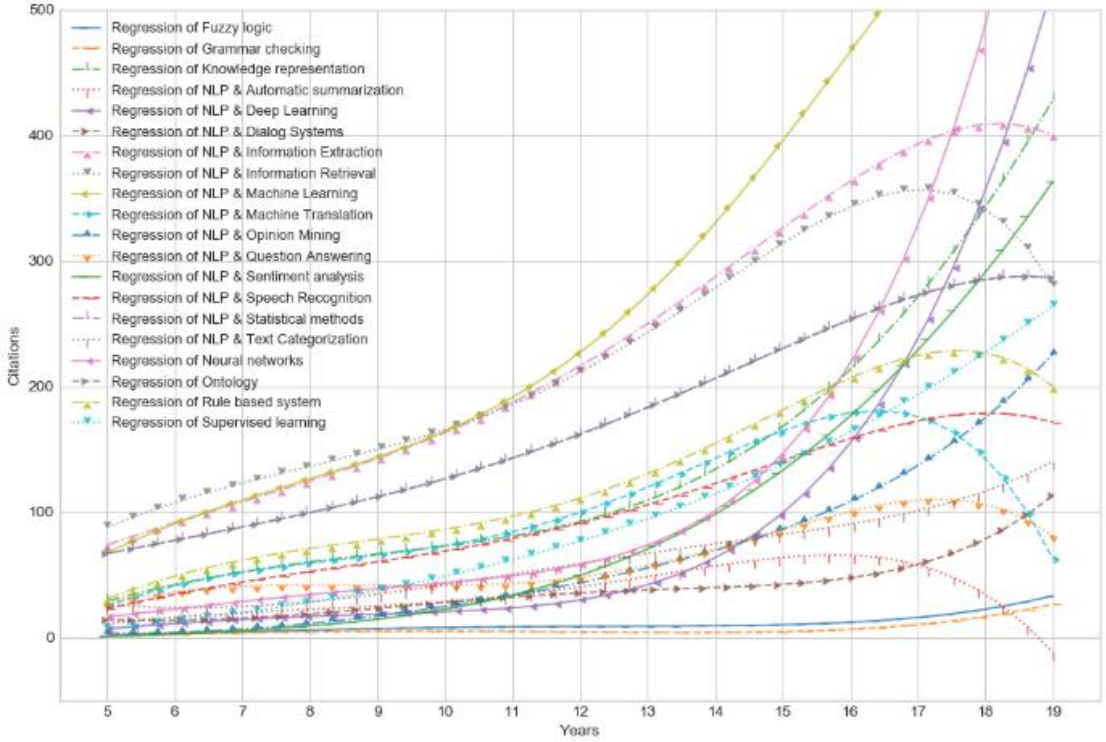

Figure 2 Publication activity Regression

Also, these categories was analyzed by citations count. Below given the graph representing the citations of each category by the year when paper was published.

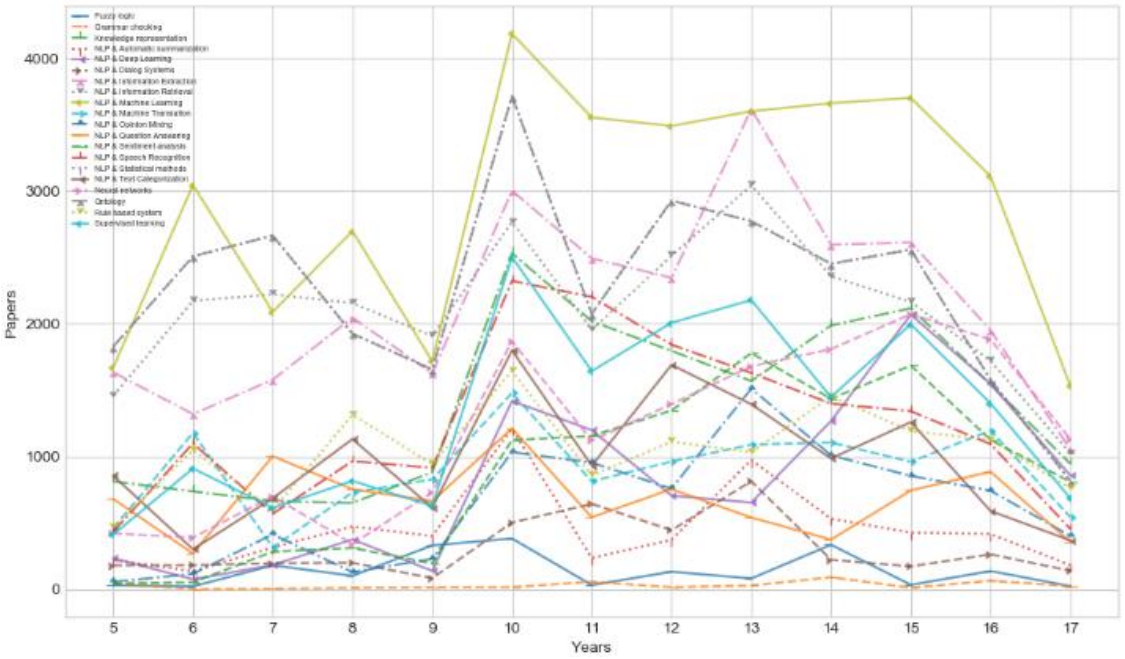

Figure 3 Citations count

It can bee seen from the graph which year's publications were the most citated i.e. most popular.

Next, D1 and D2 indicators are calculated according to the equations (1-2).

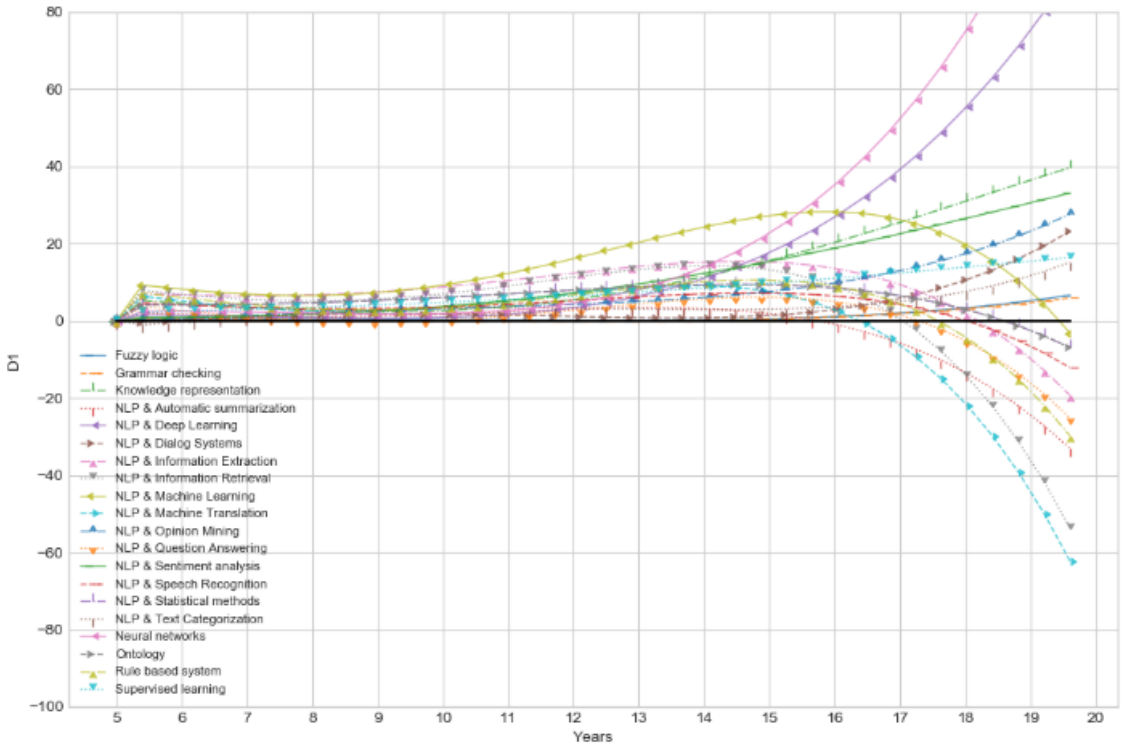

Figure 4 D1 (Speed)

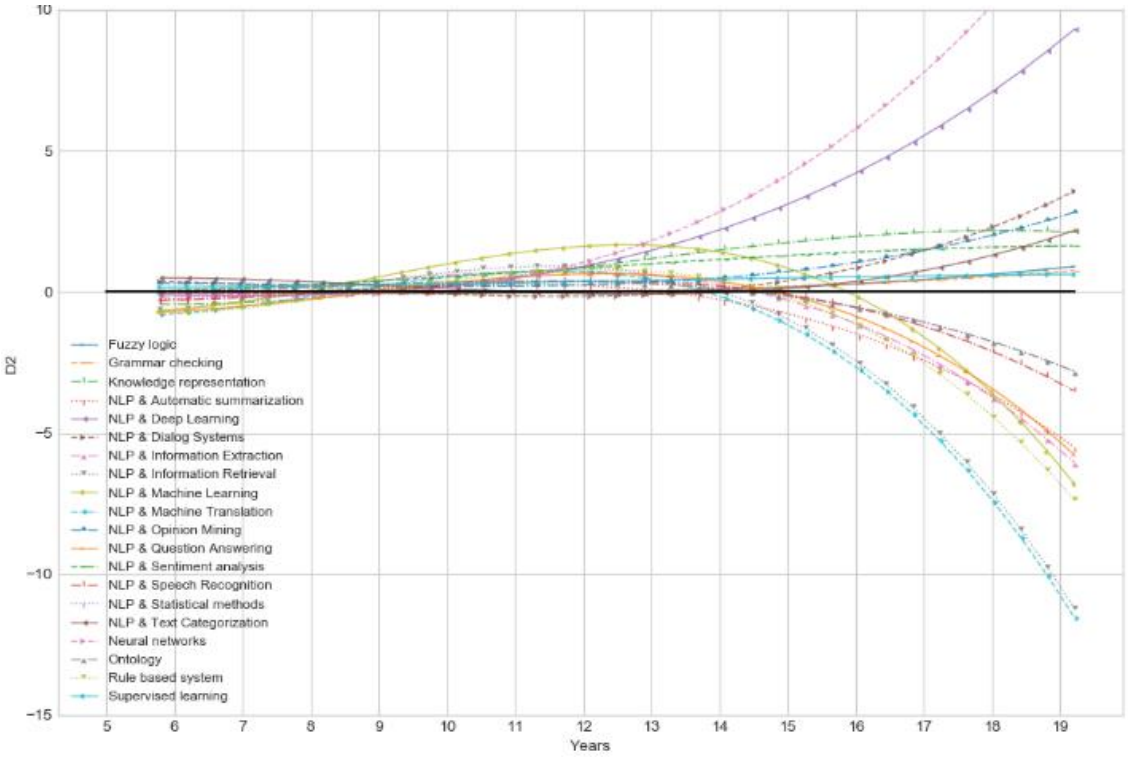

Figure 5D2 (Acceleration)

4 Conclusion

Selecting accurately regression degree and search term coefficients it is possible to analyze and explain growth in publication activity, based on bibliometric indicators, the rate of change of speed D1 and the acceleration of change D2. The positive value of D1 reflects the fact of an increase and D2 characterizes its rate in the growth of publication activity in the field of research. On the other hand, a negative valueindicates a slowdown in publication activity compared to previous periods. According to this indicators we can propose an increasing interest in domains such as NLP&DL, NLP&Opinion Mining, stability of interest in Fuzzy Logic and decrease in NLP &InformationExtraction and Supervised Learning and etc. Such scientometric indicators allows to reveal interest trends in different domains of research from objective point of view.

Acknowledgments

The authors are grateful to Dr. Denis Kosyakov (The state public scientific technological library of Siberian branch of the Russian Academy) who was kindly provide data for calculations.

The work was funded by grant No BR05236839 of the Ministry of Education and Science of the Republic of Kazakhstan.

References

[1] Barakhnin V, Duisenbayeva A, Kozhemyakina O, Yergaliyev Y, Muhamedyev R 2018 The automatic processing of the texts in natural language. Some bibliometric indicators of the current state of this research area *Journal of Physics: Conference Series* **1117** 012001. 10.1088/1742-6596/1117/1/012001

[2] Muhamedyev R, Aliguliyev R, Shokishalov Z M, Mustakayev R R 2018 New bibliometric indicators for prospectivity estimation of research fields *Annals of Library and Information Studies* **65** 62-9

[3] Mukhamedyev R I, Kuchin Y, Deni, K, Murzakhmetov S, Symagulov A, Yakunin K 2019 Assessment of the Dynamics of Publication Activity in the Field of Natural Language Processing and Deep Learning *In International Conference on Digital Transformation and Global Society* 744-53 Springer, Cham.