STI 2018 Leiden

23rd International Conference on Science and Technology Indicators
"Science, Technology and Innovation Indicators in Transition"

STI 2018 Conference Proceedings

Proceedings of the 23rd International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Chair of the Conference

Paul Wouters

Scientific Editors

Rodrigo Costas
Thomas Franssen
Alfredo Yegros-Yegros

Layout

Andrea Reyes Elizondo
Suze van der Luijt-Jansen

The articles of this collection can be accessed at https://hdl.handle.net/1887/64521

ISBN: 978-90-9031204-0

23rd International Conference on Science and Technology Indicators (STI 2018)
"Science, Technology and Innovation indicators in transition"
12 - 14 September 2018 | Leiden, The Netherlands
#STI18LDN

Stochastic analysis of citation time series of emergent research topics[1]

Maximilian Förster[*], Birgit Stelzer[*], Edgar Schiebel[**]

[*]*maximilian.foerster@uni-ulm.de, birgit.stelzer@uni-ulm.de*
Institute of Technology and Process Management, University of Ulm, Helmholtzstraße 22, 89081 Ulm, (Germany)

[**]*edgar.schiebel@ait.ac.at*
Center for Innovation Systems & Policy, AIT Austrian Institute of Technology GmbH, Giefinggasse 4, Vienna, 1210 (Austria)

Introduction

Detecting and forecasting emerging research topics has become more demanded by researchers and practitioners. Bibliometrics provide a promising way to detect emerging research topics at an early stage. However, reliably forecasting the emergence of a research topic still remains a challenge. Based on the number of cited references per year of a current research topic, we used the relative knowledge growth described as time series. The time series were analyzed stochastically. As they reveal a common pattern of memory, this memory can be used to shift the relative growth factor to the future using stochastic ARMA models. An approach to forecast the emergence of a research topic using ARMA models and thus detecting emergent research topics even earlier is proposed.

Background and Motivation

Detecting and forecasting emerging research topics has become more asked not only by future researchers, but also by R&D managers and politicians wanting to find the best investments for future success. In the context of strategic foresight, emerging research topics enable to identify the most promising technologies at a very early stage. Especially in quickly evolving industries, enterprises can gain strategic advantage by identifying the next successful technology even earlier and with more reliability than their competitors. On the other hand, politicians are supposed to fund promising technologies that are valuable for the society.
The first has already been made in 1963 with the introduction of bibliographic coupling to identify and delineate research issues within a given scientific framework (Kessler, M. (1963)). Since then, methods improved strongly, for example due to enrichment with text and semantic similarities (Yau, CK. et al 2014).) or visualization of research fronts and their knowledge bases (Schiebel E. (2015)). Bibliographic coupling and co-citation analysis have been proven to be reliable methods for clustering and delineating research topics (Boyack, K.W. & Klavans, R. (2010).).
According to Rotolo et al. (2015), emergent technologies show a radical novelty with a potential prominent impact and coherence that persists over time. Mund and Neuhäusler

[1] This work was supported by the Austrian COMET-Program (Projects K2 XTribology and K2 acib –Austrian Centre of Industrial Biotechnolog)

(2015) found qualitative factors regarding publication behavior that indicate, depending on research disciplines, emergent topics and Jarić et al. (2013) established a relationship between the age of references and the publication rate within a respective research field. Quantitative methods for the identification of emergent research topics are for instance proposed by Small, Boyak and Klavans (2014) with citation and co-citation analysis based on growth and newness, Huang et al. (2017) by tracing technology pathways based on co-classification and co-word analysis or Schiebel and Asenbeck (2017) with the Knowledge Growth Factor (KGF). The latter quantifies, based on the publications' references of a research topic, its knowledge growth and thus compares the emergence of research topics within a research field at a certain point of time according to the definition of a relatively fast growth. However, none of these measures provide a time series of emergence.

Bildosola et al. (2017) got a step further and established time series for a monthly forecast of emerging research topics. They followed a stochastic approach based on research activity of emerging research topics including trends, cycles and seasonal as well as irregular components. However, this method is lacking an interpretable measure of emergence.

We want to close this research gap and provide an interpretable and comparable measure of emergence for research topics being described over time. Therefore, we develop a measure of emergence based on the definition of Rotolo et al. (2015): A research topic is emergent if it shows a relatively fast knowledge growth. Since knowledge growth is based on previous generated knowledge, the development of an emerging research topic over time, is supposed to have a kind of memory. For example, Shibata et al. (2008) estimate the delay, until a publication is being initially cited after being published, at one or two years. In practice, researchers notice recently published valuable work, then take their time to expand the knowledge thoroughly and submit their own research paper based on the knowledge of existing work. After publication, other researchers have access to their knowledge contribution and the cycle starts again. Therefore, very successful knowledge growth might come up in cycles, since very valuable research results are more likely to create again valuable knowledge.

Since we propose a measure for emergence being described over time, we can give insight in the so-called memory of emergent research topics.

That leads to following research questions: Does knowledge growth of emergent research topics have a common pattern of memory? Is it possible to shift an existing memory to the future o get a more detailed forecast?

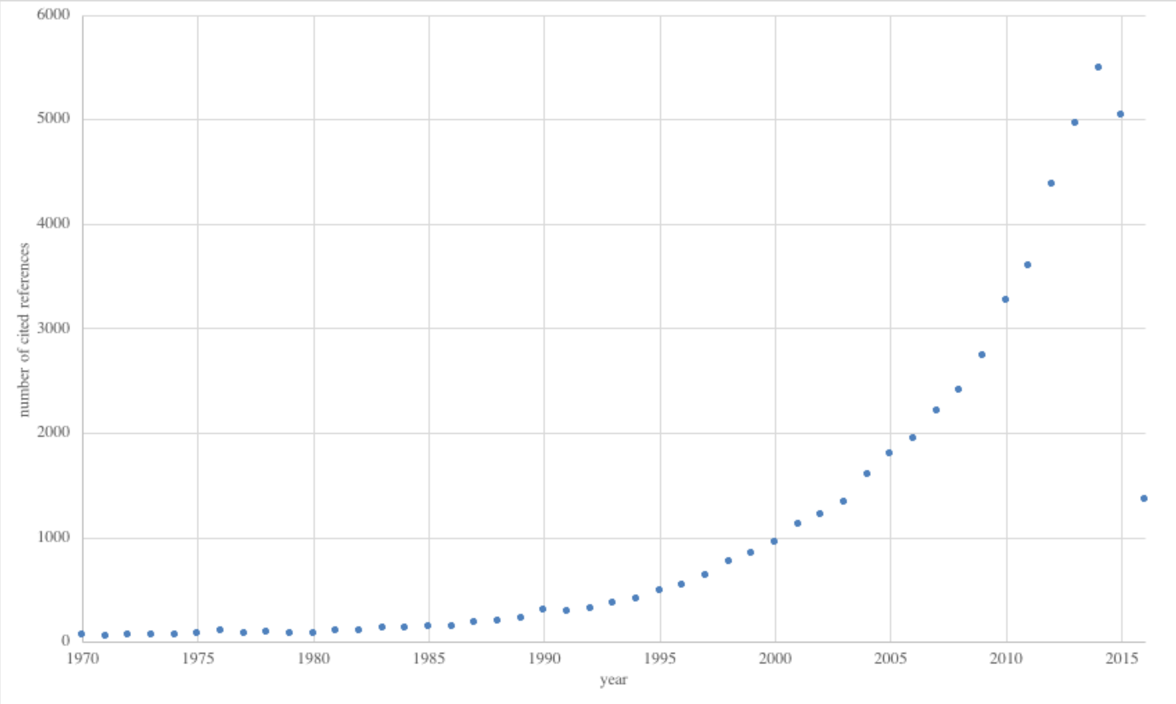
First, we introduce the relative knowledge growth over time. The presentation of the auto regression as an indicator for memory of time series follows as well as the ARMA model.

Methods and Data

The knowledge base of an existing scientific work is located at its references. Out of this knowledge base, the number of cited references per year represents the knowledge growth in the respective year. According to Rotolo, D. et al (2015) emergent research topics are growing faster than non-emergent research topics and should also have a higher knowledge growth per time unit than non-emergent research topics of the same size and discipline.

The knowledge growth per year is illustrated as a citation time series in Fig 1 in the research field of biotechnology in the year 2016. The curve shows a typical exponential growth except for the values in the years 2015 and 2016. The exponential growth can empirically be observed for citation time series of all research topics (Schiebel. E. & Asenbeck, B. (2017)) because researchers are supposed to cite the most recently published references. The significantly lower values in the years 2016 and 2017 are caused by the delay before new publications are firstly cited in other publications. Shibata *et al.* estimated this delay to be approximately one or two years, which is consistent with declining values for the last two years.

Fig 1: Number of cited references per year as a knowledge base of 4899 publications in the research field of biotechnology in the year 2016.



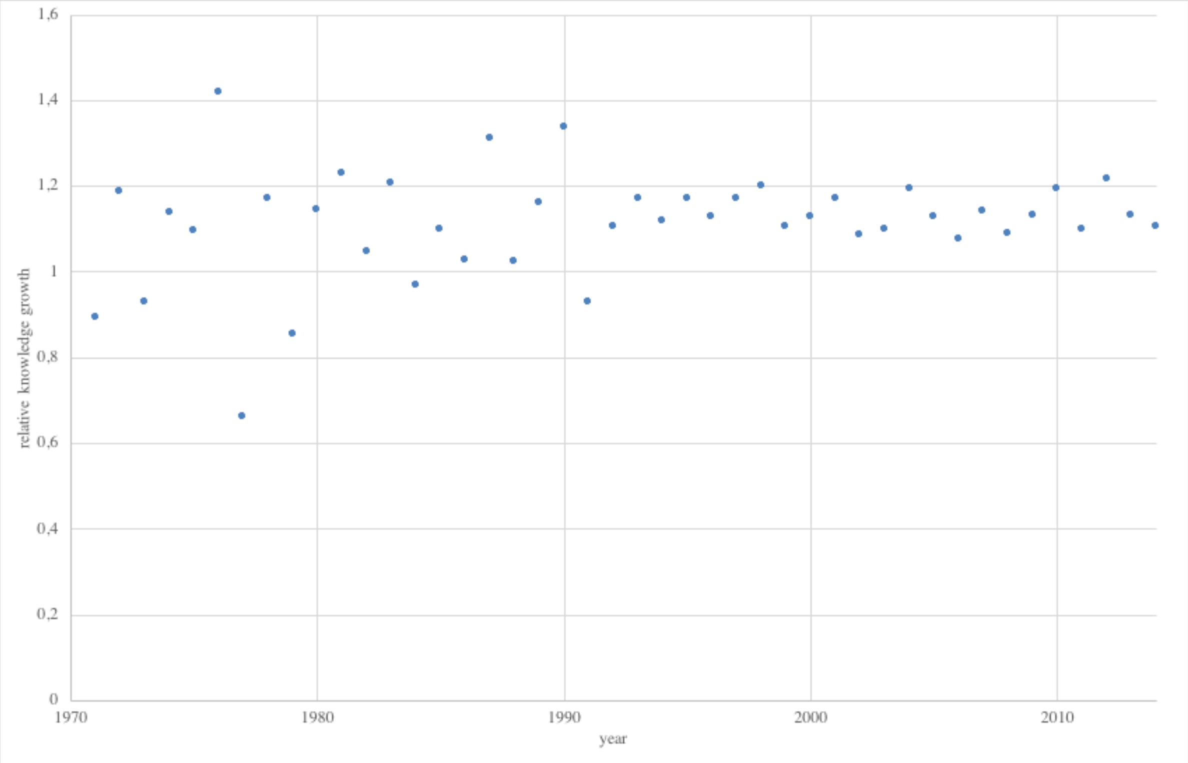
The relative knowledge growth as a measure for emergence

We introduce the relative knowledge growth since their time series can start at different years. We define the following growth process with *t* as a discrete parameter counting years and *t=0* referring to the first year of observation:

$$X_t = a * X_{t-1} \qquad (1)$$

For a smooth exponential growing time series, the growth rate *a* is a constant. We assume that *a* as the relative knowledge growth of a research topic is a measure of emergence. Fig 2 shows as an example for the time series of Fig 1. In contrast to Fig 1, the measure of emergence can be directly seen in the time series of Fig 2. So far, this is according to the definition of emergence by Rotolo *et al.* (2015).

Fig 2: Relative knowledge growth per year for the knowledge base of 4899 publications in the research field of biotechnology in the year 2016.



Furthermore, Fig 2 shows that the relative knowledge growth is somehow spread and distributed around a constant value. The variance of the relative knowledge growth is due to the stochastic components of absolute knowledge growth (Bildosola, I., Gonzalez, P., Moral, P. (2017)) which are caused by more successful or less successful years of knowledge contribution to the research topic.

Autocorrelation as an indicator for memory of time series

The spread values in Fig 2 suggest, that the time series might have a memory. The memory of a time series can be detected by its autocorrelation function. The autocorrelation ρ_τ according to the time interval τ of a time series indicates, in which form and strength values at a distance of τ are correlated. The autocorrelation takes on a value between -1 (fully anticorrelated) and +1 (fully correlated). The autocorrelation function maps all available autocorrelations ρ_τ against their time intervals τ. The autocorrelation ρ_τ of real data sets – which are in our case especially time series of relative knowledge growth – is calculated by the sample autocorrelation with T as the number of time steps and $\bar{x}$ as the mean of the sample:

$$\rho_\tau = \frac{\sum_{i=1}^{T-\tau}(x_{i+\tau}-\bar{x})(x_i-\bar{x})}{\sum_{i=1}^{T}(x_i-\bar{x})^2} \qquad (2)$$

For real time series, the dependence between values implied by a correlation is only meaningful when backward-oriented. A correlation implied by a correlation according to the

time interval τ provides information on the dependence of the values at time $t+\tau$ to the values at time τ. Not the other way around.
While the autocorrelation for the white noise process already drops almost perfectly to 0 for a time interval of 1, the memory of the random walk process is clearly visible in the autocorrelation. The autocorrelation function decreases very slowly and is close to 1 for small time intervals. Both stochastic processes are special cases of ARMA processes which are discussed in the following section.

ARMA model for description of time series with memory

When modeling and forecasting time series, a pattern of memory can be considered by a stochastic ARMA process. The ARMA model describes the evolution of a random variable over time, depending on past values. It can therefore model stochastic processes with a memory which means stochastic processes with a non-zero autocorrelation function. An ARMA (p, q) process is characterized by following equation:

$$\boldsymbol{X_t = \varepsilon_t + a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p} + b_1 \varepsilon_{t-1} + b_2 \varepsilon_{t-2} + \cdots + b_q \varepsilon_{t-q}} \quad (3)$$

The parameters a_i and b_i can be chosen. The variable ε_t denotes a Gaussian distribution with fixed mean value and fixed variance. Both the mean value and the variance can be chosen too. It contains p past values and q past random numbers.
Prerequisite for the applicability of ARMA models to real data are Gaussian distributed random variables with a constant variance. In practice, the length of the memory is read off at the autocorrelation function of the data. Subsequently, the parameters a_i and b_i as well as the variance and the mean value of the Gaussian distribution of the ARMA (p, q) models can be determined by statistical data fits.

Data

The theoretical considerations were tested on time series of cited references per year which were extracted by two different datasets: publications from the research field of biotechnology in 2016 and publications from the research field tribological wear in 2015. The time series of relative knowledge growth for each research topic was generated in four steps:
1. Detection of research topics: The research topics of biotechnology were determined on the basis of 4899 publications from the year 2016, the tribology on the basis of 2033 publications from the year 2015 via bibliographic coupling according to (Schiebel E. (2015)).
2. Generation of time series of cited references for each research topic: the reference list including publication years of cited references of each publication were available. For each research topic, a time series was generated.
3. Generation of time series of relative knowledge growth for each research topic according to formula 2.
4.Clearance of data: The observed period of relative knowledge growth was restricted. The time series of references in the biotechnology research field spanned 15 years (2000-2014), as did the time series of references in the tribological wear research field (1999-2013). See Tables 3 and 4 in the appendix for the number of references and the median of relative knowledge growth of each research topic in the period considered after data clearance.

Results

For all research topics of the two data sets, both the median of the relative knowledge growth time series and the GINI index as a measure of emergence from Schiebel and Asenbeck 2017 was calculated. The median of the relative knowledge growth strongly correlates with the

GINI index for the observed datasets and is statistically significant for the respective sample sizes for both the biotechnology research field and the tribology research field according to a two-sided t-test.

For both observed datasets with research topics in biotechnology and tribological wear, the autocorrelations at small time intervals are strikingly different from 0. Thus, the relative knowledge growth does not follow a purely randomly distributed process such as the white noise process. In fact, two characteristic features can be observed in the autocorrelation functions of the observed research topics:

1. The autocorrelation initially falls into negative for the time interval $\tau=1$. After a year with an above-average amount of valuable knowledge contribution for the knowledge base of current scientific work, a year with a below-average knowledge gain is more likely to follow and vice versa.
2. The autocorrelation increases after the lowest point again and usually has a maximum for a time interval between two and four years.

Successful knowledge contributions in science repeat after two years at the earliest. This is consistent with the fact that scientific work is quoted only one or two years after publication. The form of autocorrelations does not differentiate between emergent or non-emergent research topics. Fig 3 and 4 show two autocorrelation functions for the same time series of biotechnology research fields as in the section above. Fig 3 corresponds to the highly emergent research topic “Mass spectrometry” with a success cycle of 3 and Fig 4 corresponds to the non-emergent research topic “Integrated and Continuous Processing of Recombinant Proteins” with a success cycle of 4.

Fig 3: Autocorrelation function for the highly emergent research topic “Mass spectrometry” with a median of 1,45 and a success cycle of 3. Number of total cited references: 521.

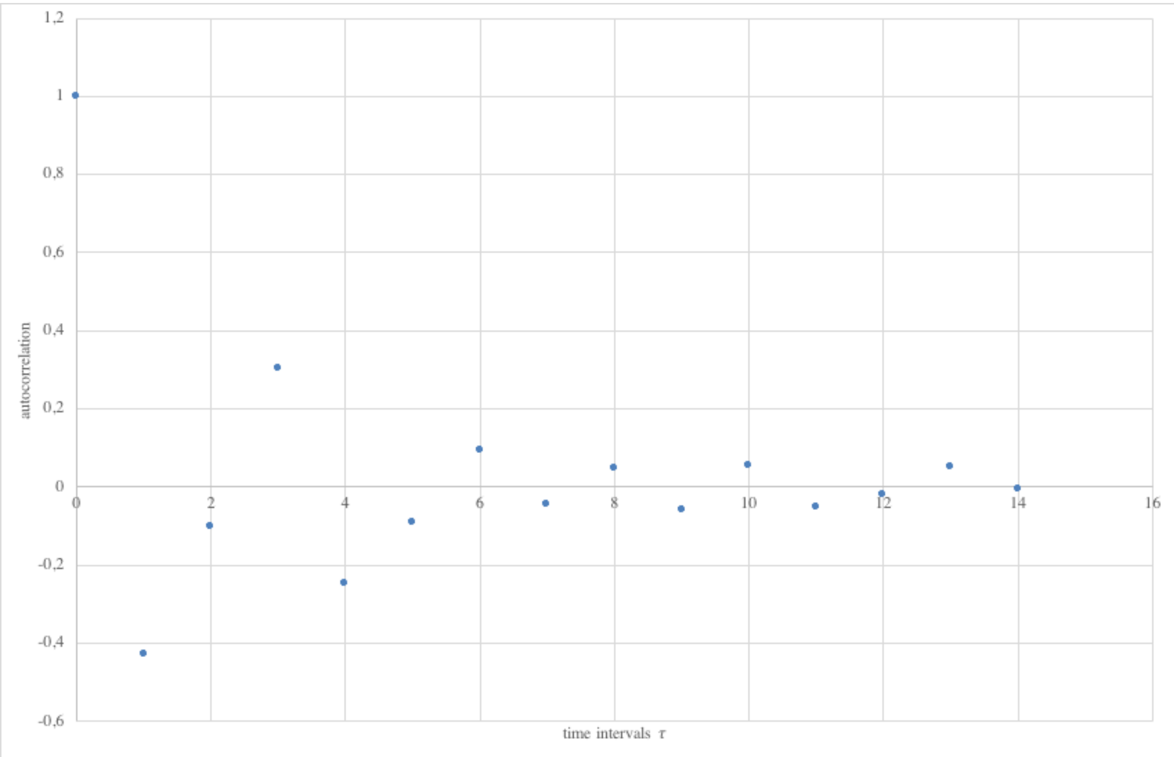
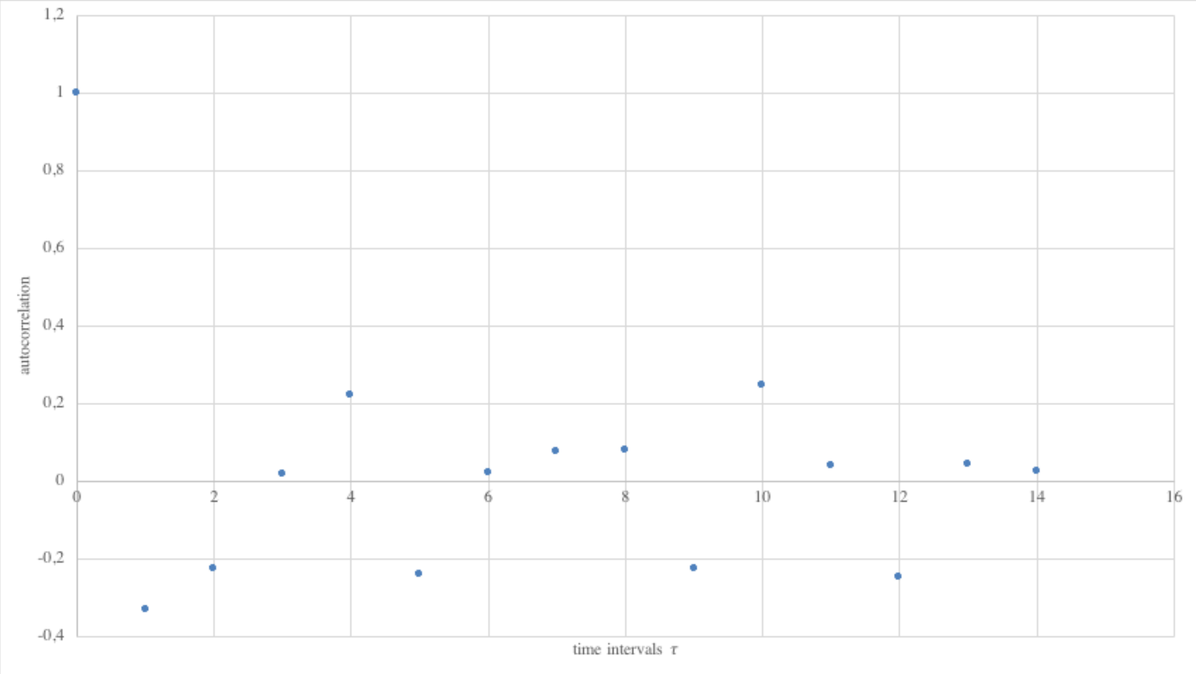


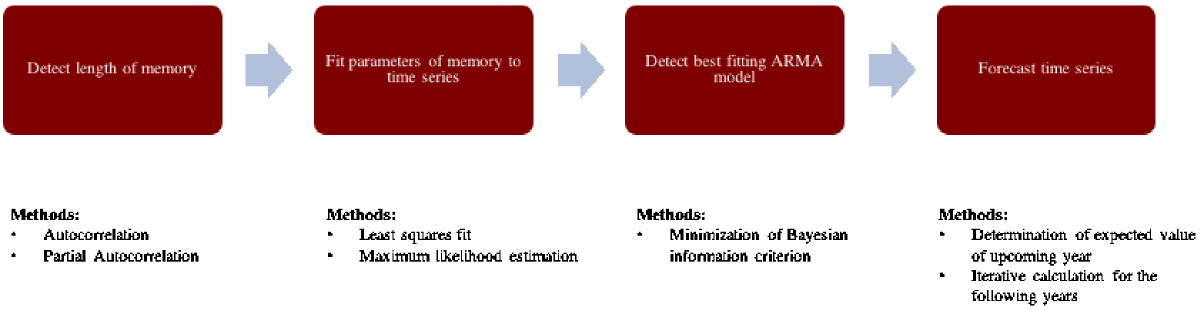
Fig 4: Autocorrelation function for the non-emergent research topic “Integrated and Continous Processing of Recombinant Proteins” with a median of 1,10 and a success cycle of 4. Number of total cited references: 781.



Approach to shift the memory of time series of relative knowledge growth to the future
As the time series of the relative knowledge growth have a common pattern of memory for each observed research topic, this memory could be shifted to the future for a more reliable and more detailed forecast of emergence and thus enable to detect emergent research topics even earlier than simply assuming the median of relative knowledge growth or a linear trend curve of the relative knowledge growth as a measure of emergence.
Summarizing this approach: The memory of today is extrapolated into the future with a stochastic process. However, using the ARMA model as a stochastic process with memory, only the length of memory of the real data should be assumed for the model. Furthermore, the time horizon of the forecast shouldn’t exceed the length of memory of the real data.

Given that the two prerequisites Gaussian distributed random variables and a constant variance are fulfilled, an ARMA process can be fitted to the concerning time series of the research topic. Both the number of parameters and the value of the parameters of ARMA models can be individually adapted to the time series of any research topic. Different research topics have different positions of the maxima in the autocorrelation function, therefore memories of different lengths and a different number of parameters to fit. Following the Box-Jenkins method (Box, G.E.P., Jenkins, G.M. (1994)) as a method to fit an ARMA process on real datasets, an illustrative approach is proposed in Fig 5 using ARMA models to forecast the relative knowledge growth of a research topic as a measure of emergence at an early stage of emerging research topics.

Fig 5: Illustrative approach to forecast the emergence of a research topic using an ARMA process for the time series of relative knowledge growth.



Confounding factors affecting the forecast are mainly founded by a potentially lacking reliability of the data. Citation time series of research topics may be biased by personnel incentives of researchers and editors which are driven by competition and limited journal space (Fon *et al.* (2017)). For instance, Fowler *et al.* (2007) found that authors self-citations pay due to increased citations receiving from others. Also inflating the journal impact factor may lead to journal self-citations (Larivière *et al.* (2018)). In this sense, manipulations can cause partially invalid knowledge bases and citation series. Regarding the forecast of emergence, especially differences in manipulative citation manners across research topics distort the results. Since the proportion of journal self-citations vary by more than 15% across disciplines (Larivière *et al.* (2018)), we recommend to compare the relative knowledge growth of a research topic as a measure for emergence with other research topics from the same discipline. For instance, statements like "The emergent research topic X grows x% faster than average research in the research field" can be made.

Conclusions

Detecting and forecasting emergent research topics is an important research field for both science and economy. This paper uses the relative knowledge growth as a measure for emergence that is expressed as time series and thus can help to detect emergent research topics even earlier.

The relative knowledge growth is based on the knowledge base of current scientific work which means the number of references of current publications. It expresses how much knowledge was contributed to a research topic in a certain year compared to the previous year. The higher the relative knowledge growth, which usually fluctuates around a constant, the faster the evolvement of the concerning research topic. Based on the definition of a relative fast growth, the relative knowledge growth serves as a measure of emergence.

Time series for different research topics show a common pattern of memory, observed with the autocorrelation function, which can be shifted to the future for a better forecast. Therefore, this paper proposes an approach to forecast the emergence of a research topic which considers the memory of the time series by modelling it with a stochastic ARMA process. Further research is needed for a more automated way to apply ARMA models for a forecast of emergence.

Furthermore, the relative knowledge growth has been verified as a measure of emergence only for two datasets of the research fields biotechnology and tribological wear. These two

research fields were chosen because in principle they show a different citation behavior. However, for a statistical validation, the relative knowledge growth as a measure for emergence should be tested on a more significant amount of data. The pattern of memory should also be verified with other research fields and, if different, be interpreted. Although the two chosen datasets are very different and still show a common pattern of memory, this is not sufficient for a generalization.
Other approaches beyond the autocorrelation function might sharpen the understanding of the memory of knowledge bases of emergent research topics. A Fourier analysis of the time series of the relative knowledge growth might give a valuable insight in the cycles of successful.
Our paper contributes to a better understanding of detecting and forecasting emergent research topics and stimulates further research in this exciting field.

References and Citations

Bildosola, I., Gonzalez, P., Moral, P. (2017). An approach for modelling and forecasting research activity related to an emerging technology. *Scientometrics.* 112(1) 557–72.

Box, G.E.P., Jenkins, G.M. (1994).Time Series Analysis-Forecasting and Control. *J Mark Res. 2.* 14(2), 199–201.

Boyack, K.W. & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society of Information Science and Technology* 61(12), 2389–404.

Fong, E. A., Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PloS one*, 12(12), e0187394.

Fowler, J., Aksnes, D. (2007). Does self-citation pay? *Scientometrics.* 72 (3), 427-437.

Huang, Y., Zhu, D., Qian, Y., Zhang, Y., Porter, A.L., Liu, Y. (2017). A hybrid method to trace technology evolution pathways: a case study of 3D printing. *Scientometrics*, 111(1), 185–204.

Jarić, I., Knežević-Jarić, J., Lenhardt, M. (2013). Relative age of references as a tool to identify emerging research fields with an application to the field of ecology and environmental sciences. Scientometrics, 100(2), 519–29.

Kessler, M. (1963). Bibliographic coupling between scientific papers. *Journal of the American Society for Information Science and Technology*, 14(1), 10-25.

Larivière, V., Cassidy R. (2018). The Journal Impact Factor: A brief history, critique, and discussion of adverse effects. *arXiv preprint arXiv:1801.08992*.

Mund, C., Neuhäusler, P. (2015). Towards an early-stage identification of emerging topics in science—The usability of bibliometric characteristics. *Journal or Informetrics*, 9(4), 1018–33.

Rotolo, D., Hicks, D., Martin, BR. (2015). What Is an Emerging Technology? *Research Policy*, 44(10), 1827–43.

Schiebel E. (2015) Mapping the Spreading of Cited References over Research Fronts of Bibliographically Coupled Publications. *Reinventing Information Science Networked Soc Proc 14th International Symposium of Information Science, Zadar*, 66, 404–9.

Schiebel. E. & Asenbeck, B. (2017). The Knowledge Growth Factor KGF as a new indicator for the quantification of the emergence of research issues - The case of tribological wear. *Atlanta Conference on Science and Innovation Policy*

Schiebel. E. & Asenbeck, B. (2017). The Knowledge Growth Factor - An indicator for the quantification of the emergence of new research issues. *Technological Forecasting and Social Change*, submitted.

Shibata, N., Kajikawa, Y., Takeda, Y., Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*. 28(11), 758–75.

Small, H., Boyack, K.W., Klavans, R. ((2014). Identifying emerging topics in science and technology. *Research Policy*, 43(8), 1450–67.

Shibata, N., Kajikawa, Y., Takeda, Y., Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*. 28(11), 758–75.

Yau, CK., Porter, A., Newman, N, Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, 100(3), 767–86.

Appendix

Table 1: Correlations and significances of the median and the mean of relative knowledge growth with the GINI index as a measure of emergence.

Research field	**Biotechnology**	**Tribological wear**
Correlation median of relative knowledge growth with GINI index	0,68	0,72
Significance of the correlation median of relative knowledge growth with GINI index	There is a 99% secure statistical correlation.	There is a 99% secure statistical correlation.
Correlation mean of relative knowledge growth with GINI index	0,13	0,30
Significance of the correlation mean of relative knowledge growth with GINI index	There is no secure statistical correlation.	There is no secure statistical correlation.

Table 2: Overview of the database.

Research field	**Biotechnology**	**Tribological wear**
Publication year	2016	2015
Number of publications	4.899	2.033
Search process	Search for all publications with publication year 2016 in the Web of Science database, that cite publications of the biotechnology research organization ACIB at least once between 2011-2017.	Search for all publications with publication year 2015 in the Web of Science database, that have „tribological wear“ or similar words as keywords.
Initial time span of cited references	1970-2016	1990-2015
Total number of cited references in initial time span	79561	60076
Number of research topics	22 resarch topics + one time series including cited references of all publications oft the research field as a reference.	24 resarch topics + one time series including cited references of all publications oft the research field as a reference.
Comment	The selected publications do not completely cover the research field of biotechnology due to the search process. They form part of the research in biotechnology in 2016. However, this does not affect the analyzes.	Two research topics were excluded from the analysis because their time series had too less data points (17 and 49 cited references in total in initial time span).

Table 3: Research fronts of the biotechnology research field with sum of references and median of relative knowledge growth in the period considered (2000-2014).

Research topic	**Sum of references**	**Median of relative knowledge growth**
Inducing pluripotent stem cells by reprogramming somatic cells	1326	1,23
Nitrate transport, uptake and regulation in plants	1143	1,06
Biofuel cells, biosensors and bioelectrocatalysis due to electron transfer	753	1,26
Trichoderma reesei – lignocellulose degration and source of cellulase	1538	1,13
Lignin depolymerization by white rot fungi	704	1,17
DNA barcoding of plant pathogenic fungi	893	1,13
Root microbiota	1934	1,19
Trimeric autotransporter adhesins (TAAs)	161	1,00
Mass spectrometry (MS)-based proteomics	521	1,45
M5C RNA methylome	671	1,13
Analysing lipids	312	1,18
Producing bone graft substitutes in flow perfusion bioreactors	264	1,21
Lipid particles	866	1,07
Hydrolysis of poly(ethylene terephthalate)	752	0,96
Itaconic acid production	211	1,14
Asymetric biocatalysis with “Old Yellow Enzyme” family	603	1,11
Production of amino acids and chiral amines by transaminase reactions	1580	1,20
Microreactors for immobilized microfluidic enzymes	704	1,24
Integrated and continuous processing of recombinant proteins	781	1,10
Modelling metabolic networks	1110	1,16
Chinese hamster ovary cell lines for cell engineering	1139	1,20
Pichia pastoris as a platform for the production of proteins	2647	1,06
All publications	39061	1,13

Table 4: Research fronts of the tribological wear research field with sum of references and median of relative knowledge growth in the period considered (1999-2013).

Research topic	**Sum of references**	**Median of relative knowledge growth**
Adaptive Coatings	801	1,09
Bond strength	269	0,93
Brake pad	594	1,09
Contact and rubbing of flat surfaces	587	0,96
Diamond like carbon (DLC)	2297	1,01
Electrodeposition technique	552	1,11
Electrostatic separation	575	1,16
Graphene	1411	1,30

Hard and multilayer coatings	2167	1,08
High velocity oxygen fuel (HVOF)	968	1,03
Ionic liquids	1120	1,28
Lubricant additives	698	1,10
Luminescent materials	525	0,95
Nanoparticles	1029	1,10
Oral texture and sensory research	230	1,17
Plasma electrolytic oxidation	401	1,15
Polymer matrix and hybrid composites	180	1,00
Pumping pressure	19	1,00
Superhard coatings and sensing indentation	1016	1,12
Surface texturing to improve sliding	2045	1,16
Thermoelastic instabilities	57	1,00
Triboelectric nanogenerator	6678	1,33
Water lubrication	744	1,02
Wear behaviour biomedical alloys and corrosion	1002	1,12
Zink and oil behaviour	513	1,06
All publications	26478	1,15

Table 5: Tested measures of emergence based on the time series of relative knowledge growth of a research topic.

Tested measures of emergence	**Statistical proved by comparison with GINI index**
Median of relative knowledge growth	Yes
Average of relative knowledge growth	No
Strenth of memory parameterized by the accumulated absolute values of the autocorrelations up to a certain time interval s	No
Y-intersept of a linear regression for the time series of relative knowledge growth	No
Slope of a linear regression for the time series of relative knowledge growth	No