DOI. 10.1109/ISETC56213.2022.10010222

Evaluating Research Trends Using Key Term Occurrences and Multivariate Mann-Kendall Test

Christian-Daniel Curiac
Department of Computer and Information Technology
Politehnica University of Timisoara
Timisoara, Romania
christian.curiac@cs.upt.ro

Mihai Micea
Department of Computer and Information Technology
Politehnica University of Timisoara
Timisoara, Romania
mihai.micea@cs.upt.ro

Abstract—**This paper offers a quantifiable means for assessing the trend in the interest of scientific community in a particular research theme. Representing the research theme as a set of key terms, we investigate its trend within the scientific literature, using a corpus of paper metadata belonging to flagship journals and conferences that are relevant for the encompassing domain. Our method employs a multivariate version of the Mann-Kendall test that is applied to a multivariate time series of key term occurrences in order to discover if the trend of the research theme is ascending, descending or if there is no monotonic trend. Finally, an illustrative case study, examining the effectiveness of the proposed method for three research themes from Electronic Design Automation field, is presented.**

Index Terms—**research theme, Mann-Kendall test, multivariate time series, term occurrence, natural language processing**

I. Introduction

Bibliometric databases are an important source of insights for scientific community. They provide meaningful information to evaluate diverse research facets including its relevance, impact and performance or to discover trends or hot spots. In the attempt to automate these types of activities, a branch of artificial intelligence, namely Natural Language Processing (NLP), provides the much needed means to examine, understand, and extract meaning from text corpora.

Researchers are recurrently faced with the question of whether their work on a specific topic is still in trend or they need to start working on another, more promising theme. The answer is neither simple nor immediate needing a lot of tedious work to survey the abundant up-to-date literature and draw appropriate conclusions. This reason, coupled with the needs to correlate the research with the advancements originating from emergent domains and to reduce the subjectivity in research theme framing process, sustains the necessity of automatic NLP-based procedures to assess the research themes trends.

During the last decade, natural learning processing methods have advanced at fast pace and are now able to provide the necessary means to systematically and quickly analyze bibliometric and scientometric data to reveal research trends. In this context, an almost traditional approach to examine research trends is to analyze the time evolution of individual key terms occurrences in paper metadata records [1]. Generally, such methods process the paper metadata fields, namely title, keywords and abstract, in a bag-of-words or bag-of-entities fashion and form time series of term occurrence counts which are later categorized as "hot" or "cold" using classic trend detection tests. As a representative example, Marrone [2] used paper titles and abstracts, preprocessed as bag-of-entities, and utilized a mixture of trend indicators provided by Mann–Kendall test, Sen's slope estimation and Kleinberg burst detectiont to obtain terms' trends. A similar method was developed by Marchini et al. [3] to assess the urologic research trends previously associated to twelve key terms.

While the mentioned traditional approach considers that each research topic is associated with a single key term, our proposed methodology is based on the premise that a research theme can be more accurately formalized as a finite set of key terms. By this, instead of analyzing the trends within univariate time series we need to adopt a multivariate time series trend detection mechanism.

This paper aims to provide an effective procedure to examine the trend of a specified research theme based on the interest of scientific community encapsulated in bibliometric databases. Our method utilizes a multivariate version of Mann-Kendall trend test to identify whether a monotonic (i.e. ascending or descending) trend is present or not.

The rest of the paper is organized as follows. Section II presents the problem description and formulates the proposed NLP-based methodology to identify the research themes trends. Section III briefly reviews the Mann-Kendall trend test for univariate and multivariate time series, while Section IV provides an exemplifying case study. Finally, conclusions are drawn in Section V.

II. Problem Description and Proposed Methodology

In a NLP context, each research theme RT can be roughly modeled by a finite set of carefully selected key terms KT_q, with $q = 1, ..., Q$. We aim to develop a method to derive the trend of the given research theme RT from bibliometric information.

Choosing the set of key terms KT_q can be facilitated by identifying the key terms that are used in the particular scientific domain. This can be done using a classic NLP procedure by completing the following steps: i) identifying a list of publications that are specific and relevant to the domain; ii) extract and normalize the key terms from the document corpus; iii) rank the key terms based on their occurrence counts in the given document corpus; iv) select first hundreds or thousands (depending on the broadness, specificity and considered granularity of the domain) of key terms to represent the scientific domain.

In order to understand and quantify the trend of a specified research theme within scientific publications, we propose the following six-phase methodology that is able to provide the trend (i.e. ascending, descending or no trend):

Phase 1: Select the domain that encompasses the research theme

The proposed methodology starts by selecting the research area that encloses the given research theme. The size of the chosen domain is at the discretion of the investigator. While on one hand, selecting a larger domain may offer a broader perspective upon theme's trend, on the other hand it increases the computational time and decreases the granularity of the problem. For example, analyzing the trends of a machine learning topic within computer engineering may have different results than within the broader electrical and electronics domain.

Phase 2: Select the domain relevant publications

The second phase is devoted to the selection of the flagship periodicals of the research domain. The highly-regarded journals and periodic/annual scientific conferences are selected based on their reputation within the global scientific community and need to have a continuity of more than ten years in order to be relevant for our goal (for a viable trend detection within a time series we need at least 5-10 time-ordered values).

Phase 3: Bibliometric data acquisition and preprocessing

In this stage the set of paper metadata, corresponding to the journals and/or conferences selected in the second phase are collected and processed in an automatic or semi-automatic manner. The aim is to obtain a short text file for each paper, named processed abstract, comprising only relevant key terms.

The data acquisition and preprocessing procedure is depicted in Fig. 1 and starts with gathering the relevant bibliographic data for each journal or conference paper (i.e. paper metadata), including title, abstract, keywords, publication year and authors using specific Application Programming Interfaces (APIs). In this respect relevant scientific databases (e.g. Clarivate Web of Science, Scopus or IEEE Xplore) provide suitable APIs. Next, a data cleansing step is performed to filter out irrelevant records from the scientific database which are retrieved among the scientific papers such as: table of contents, list of contributed authors, reviewers or editors, errata, indexes. Since such misleading records have empty abstract or keyword fields, they need to be eliminated.

In the attempt to summarize the content of a given paper as well as possible, a consolidated abstract is constructed by concatenating three of the paper metadata fields, namely title, keywords and abstract. Afterwards, the essence encapsulated inside the consolidated abstracts is extracted using a bag-of-entities (BoE) model approach. This type of approach transforms a text document, in our case the consolidated abstract, in a list of meaningful representations (i.e. mentions), by associating them with entities from Wikipedia or other knowledge base [4]. We named the BoE forms of consolidated abstracts as processed abstracts which are used in our methodology in a twofold way: i) to derive the list of domain's key terms by ranking the key terms based on the number of their occurrences in the processed abstract corpus during a given time interval; and ii) to count the key terms appearances within the corpus in a specified time interval to build the multivariate time series that characterizes the research theme time evolution.

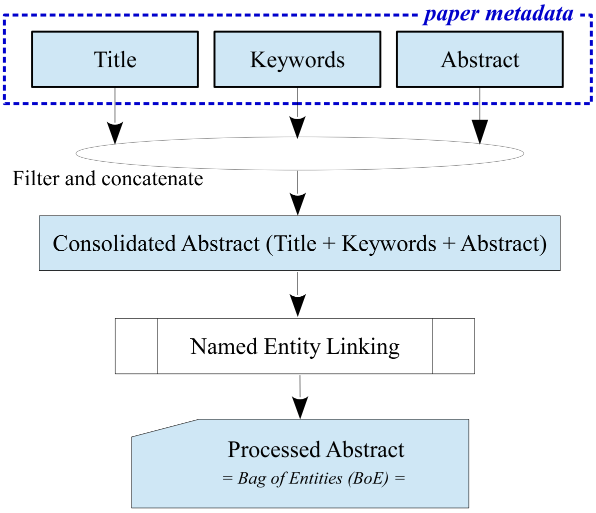

Fig. 1. Data preprocessing

Phase 4: Formalize the research theme as a finite list of key terms

Every research theme may be described using a finite set of key terms. When selecting the relevant key terms two specific rules need to be followed: i) the key terms need to be in a normalized form to reduce their randomness - if the key terms are selected from the list of domain terms obtained using the approach briefly described at the end of the description of Phase 3, this condition is already fulfilled; and, ii) the key terms need to consistently reflect every facet of the research theme including concepts, theories, methodologies, and materials.

Phase 5: Apply multivariate trend identification

We chose the multivariate version of the Mann-Kendall test to detect the overall trend of a given research theme by analyzing the multivariate time series containing the occurrences of all the key terms during the selected time period.

Phase 6: Interpreting the results

Each trend detection method offers a set of outputs (i.e. parameters) that can be analyzed and count in supporting the decisions to continue or drop the given research theme.

III. Mann-Kendall Trend Test

Mann–Kendall (MK) [5], [6] lies in the category of non-parametric statistical trend detection methods. Due to its simplicity and robustness against non-Gaussian distributed or censored data that are likely to be present in term occurrence-type time series, MK gradually becomes the standard method to detect monotonic trends in natural language processing approaches [2], [7]. In the following paragraphs, a brief summary of this method, including the univariate and multivariate case, is presented.

Considering a time-ordered sequence of N data points X_i with $i = 1, 2, \ldots, N$, the MK test examines changes in signs of the variations between consecutive values by evaluating the S-statistic according to the following formula:

$$S = \sum_{k=1}^{N-1} \sum_{j=k+1}^{N} sgn(X_j - X_k), \tag{1}$$

with $sgn(X)$ being the signum function

$$sgn(X_j - X_k) = \begin{cases} 1 & for \quad (X_j - X_k) > 0 \\ 0 & for \quad (X_j - X_k) = 0 \\ -1 & for \quad (X_j - X_k) < 0 \end{cases} \tag{2}$$

In the case of a sufficiently large number of data points N, for example $N \geq 10$, the S-statistic displays an almost normal distribution characterized by a zero mean $E(S) = 0$ and a variance $\sigma^2(S)$ provided by:

$$\sigma^2(S) = \frac{1}{18}\left[n(n-1)(2n+5) - \sum_{k=1}^{M} r_k(r_k-1)(2r_k+5)\right]. \tag{3}$$

In (3) M denotes the consecutive data points having the same value (i.e. tied groups), while r_k represents the k^{th} tied-group's rank. It is noteworthy to mention that a positive value of S indicates an upward trend for the X_i time series, while a negative value characterizes a downward trend.

Using equations (1) and (3) we may simply obtain the MK Z-statistic (Z_{MK}), with a zero mean and a unit variance, in the form:

$$Z_{MK} = \begin{cases} \frac{S-1}{\sigma(S)} & for \quad S > 0 \\ 0 & for \quad S = 0 \\ \frac{S+1}{\sigma(S)} & for \quad S < 0 \end{cases}, \tag{4}$$

where a negative value characterizes a decreasing trend, while a positive one describes an increasing trend.

Lettenmaier [8] extended the univariate MK test presented in the first part of this section, proposing a multivariate version of the Mann–Kendall method able to detect trends in a set of joint time-ordered sequences (i.e. multivatiate time series). The multivariate procedure combines the information from individual time-series and a corrected S-statistics can be obtained using the covariance matrix [9]. According to [10], [11], the following steps need to be undertaken:

i) The Mann-Kendall S-statistic scores are computed for each variate X separately using (1).

ii) The covariance matrix Γ is build according to [9], each of its elements being calculated using the formula:

$$\Gamma_{XY} = \frac{1}{3}\left[K + 4\sum_{j=1}^{N} R_{jX}R_{jY} - N\left(N+1\right)^2\right], \tag{5}$$

where X and Y are two univariate time series of length N that compose the overall multivariate time series, while K and R_jX are computed using

$$K = \sum_{1 \leq i < j \leq N} \operatorname{sgn}\left[\left(X_j - X_i\right)\left(Y_j - Y_i\right)\right] \tag{6}$$

and

$$R_{jX} = \frac{1}{2}\left[N + 1 + \sum_{i=1}^{N} \operatorname{sgn}\left(X_j - X_i\right)\right] \tag{7}$$

iii) The corrected Z-statistics for the multivariate time series is derived using the following formula:

$$Z = \frac{\sum_{i=1}^{d} S_i}{\sqrt{\sum_{j=1}^{d}\sum_{i=1}^{d}\Gamma_{ij}}}, \tag{8}$$

where d denotes the number of components (i.e., variates) of the multivariate time series, S_i are the S-statistic scores corresponding to each variate and Γ_{ij} denotes the elements of the symmetric covariance matrix Γ.

In our experiments, an effective algorithm to perform multivariate MK trend detection, implemented inside pyMannKendall Python package [12] - *multivariate_test()* - and derived from [13], is used.

IV. Illustrative Case Study

In order to exemplify our methodology, we examine the existence of monotonic trends for three research themes:

RT 1: Computational modeling of power consumption in mixed signal systems based on Monte Carlo methods;

RT 2: Design of integrated circuits for IoT applications optimized for energy efficiency by means of ML;

RT 3: Approximate computing for solving optimization problems in fault tolerant ML.

Phase 1: we chose the domain in which our investigation is applied to be Electronic Design Automation (EDA), a subfield of Electric and Electronic Engineering [14].

Phase 2: we selected a flagship journal that, in our perspective, accurately reflects the research developments of the EDA domain, namely the IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD). In this context, we aim to find how the trends of RT1, RT2 and RT3 are reflected in the time interval 2001-2020.

Phase 3: The TCAD paper metadata for the interval 2001-2020 were extracted as a Comma-Separated Values (CSV) file from IEEE Xplore using an automatic procedure implemented as a Python program around the API provided by IEEE.In this csv file each journal paper is described by the following attributes: title, abstract, keywords, author names, digital object identifier, publication year, and number of citations.

After filtering out the unrepresentative records (e.g. errata, lists of editors or reviewers, etc.), we applied the TagMe entity-linking method for the consolidated abstract to obtain the BoE representation of each journal paper. TagMe is a software tool able to discover meaningful strings (called ”spots”) in a given unstructured text that are also encountered as links in Wikipedia [15], [16]. By representing the papers as processed abstracts having a BoE form, we enable the counting of key term occurrences and, by this, the formation of the multivariate time series that characterizes the research theme evolution.

Phase 4: the three research themes are formalized as sets of key terms as follows:

- RT1: ”mixed signal”, ”computational modeling”, ”Monte Carlo”, ”convergence”.
- RT2: ”machine learning”, ”energy efficiency”, ”internet of things”, ”optimization problem ”;
- RT3: ”machine learning”, ”fault tolerant”, ”optimization problem”, ”approximate computing”.

Phase 5 and 6: the multivariate Mann-Kendall procedure is applied to the time series that describe the time evolution of the three research themes. The obtained Z-statistic proves a decreasing trend for RT1 (z=-0.67216607) and increasing trends for RT2 (z=4.33012701) and RT3 (z=1.70547360). To understand the time evolution of the interest of scientific community in the three research themes, we displayed the Z-statistic for RT1, RT2 and RT3 during the interval 2001-2020 in Fig. 2. As can be seen, RT1 changed its trend to decreasing in 2020 after seven years when its popularity has increased. RT2 and RT3 are keeping their upward trend for at least thirteen years.

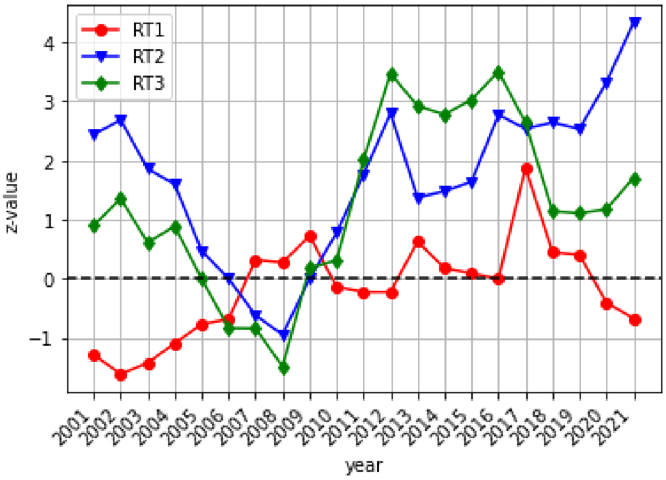

Fig. 2. Time evolution of the Z-statistic for the three research themes

V. Conclusions

This paper provides an effective NLP procedure to examine the trend of a research theme, specified as a set of key terms, based on the interest of scientific community encapsulated in paper metadata corresponding to flagship journals and conferences. The method employs a multivariate version of Mann-Kendall test, applied to a multivariate term occurrence time series, in order to identify whether a monotonic (i.e. ascending or descending) trend is present or not. This trend analysis is intended to help researchers in making a decision to drop a research theme or to find another promising research theme to work on.

References

[1] C.-D. Curiac, A. Doboli, D.-I. Curiac ”Co-Occurrence-Based Double Thresholding Method for Research Topic Identification”, Mathematics, 10(17), 3115, doi: 10.3390/math10173115.

[2] M. Marrone, ”Application of entity linking to identify research fronts and trends”, Scientometrics, 122(1), pp. 357-379, 2020, doi: 10.1007/s11192-019-03274-x.

[3] G.S. Marchini, K.V. Faria, F.L. Neto, F.C. Torricelli et al., ”Understanding urologic scientific publication patterns and general public interests on stone disease: lessons learned from big data platforms”, World journal of urology, 39(7), pp. 2767-2773, 2021, doi: 10.1007/s00345-020-03477-5.

[4] M.A. Khalid, V. Jijkoun and M.D. Rijke, ”The impact of named entity normalization on information retrieval for question answering”, In Proc. European Conference on Information Retrieval, March 2008, pp.705-710, Springer, Berlin, 10.1007/978-3-540-78646-7_83

[5] H.B. Mann, ”Nonparametric tests against trend”, Econometrica: Journal of the Econometric Society, 13, pp. 245–259, 1945, doi: 10.2307/1907187

[6] K. Kendall, ”Thin-film peeling-the elastic term”, Journal of Physics d: Applied Physics, 8(13), 1449, 1975, doi: 10.1088/0022-3727/8/13/005.

[7] C.-D. Curiac, O. Banias and M. Micea, ”Evaluating research trends from journal paper metadata, considering the research publication latency”, Mathematics, 10(2), 233, 2022, doi: 10.3390/math10020233.

[8] D.P. Lettenmaier, ”Multivariate nonparametric tests for trend in water quality”, Wiley Online Library, 1988, doi: 10.1111/j.1752-1688.1988.tb00900.x.

[9] C. Libiseller and A. Grimvall, ”Performance of partial Mann–Kendall tests for trend detection in the presence of covariates”, Environmetrics 13(1), pp.71–84, 2002, doi: 10.1002/env.507.

[10] T. Pohlert, ”Non-parametric trend tests and change-point detection”, https://CRAN.R-project.org/package=trend (accessed July 12, 2022).

[11] A. Burauskaite-Harju, A. Grimvall and C.V. Brömssen, ”A test for network-wide trends in rainfall extremes”, International journal of climatology, 32(1), 86-94, 2012, doi: 10.1002/joc.2263.

[12] M. Hussain and I. Mahmud, ”pyMannKendall: a python package for non parametric Mann Kendall family of trend tests”, Journal of Open Source Software, 4(39), 1556, 2019, doi: 10.21105/joss.01556.

[13] R.M. Hirsch, J.R. Slack and R.A. Smith, ”Techniques of trend analysis for monthly water quality data”, Water resources research, 18(1), 107-121, 1982, doi: 10.1029/WR018i001p00107.

[14] C.-D. Curiac and A. Doboli, ”Combining informetrics and trend analysis to understand past and current directions in electronic design automation”, Scientometrics, 127(10), pp. 5661-5689, 2022, doi: 10.1007/s11192-022-04481-9.

[15] P. Ferragina and U. Scaiella, ”Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)”, In Proc. of the 19th ACM international conference on Information and knowledge management, Oct. 2010, pp. 1625-1628, doi: 10.1145/1871437.1871689.

[16] P. Ferragina and U. Scaiella, ”Fast and accurate annotation of short texts with wikipedia pages”. IEEE software, 29(1), pp. 70-75, 2011, doi: 10.1109/MS.2011.122.