Research Article

Author and Keyword Bursts as Indicators for the Identification of Emerging or Dying Research Trends

Patrick Kenekayoro

Niger Delta University Address Mathematics / Computer Science Department Niger Delta University PMB, Amassoma, Bayelsa, NIGERIA.

ABSTRACT

Context: Identifying emerging research fronts is critical as it aids policy makers or funding agencies on their decisions in research policies, it is also a useful tool for guiding young researchers' research direction. Studies have successfully used techniques such as co-citation and co-word analysis of data retrieved from ISI databases to investigate emerging or dying research trends. Aim: With the advent of publicly available preprint databases such as BioRxiv, it becomes necessary to investigate if the state of the art techniques used to identify emerging research areas can be transferred to a dataset extracted from these public archives. Methods and Materials: A cluster analysis of keyword burst and author burst from data extracted from BioRxiv is used to investigate the suitability of BioRxiv dataset for the investigation of emerging or dying research trends. Results: The results showed that although the data retrieved from BioRxiv may not yet be mature enough for reliable analyses, the increased awareness shown in the exponential growth in preprint submissions suggests that this data source will be a valuable resource and the techniques described is this research can be used to discover interesting trends from preprint databases on emerging or dying research fronts.

Keywords: Biorxiv, Emerging research, Clustering, Keyword extraction, Mixed indicators model.

Correspondence
Patrick Kenekayoro
Niger Delta University Address
Mathematics / Computer Science
Department Niger Delta University PMB
581 Amassoma, Bayelsa, NIGERIA.
Email: patrick.kenekayoro@outlook.com

Received: 03-05-2020
Revised: 03-07-2020
Accepted: 19-07-2020
DOI: 10.5530/jscires.9.2.15

INTRODUCTION

Identifying emerging research areas is necessary for new researchers or PhD students when choosing topics for research projects. Funding bodies whose main objective is supporting new and promising research trends also need to effectively identify emerging and/or dying research areas in other to aid their research policies.

A number of studies have investigated techniques to identify these emerging or dying research areas.[1-6] Guo, Weingart and Börner introduced the mixed indicators model which suggests that emerging areas attract new authors and are interdisciplinary.[7]

Guo, Weingart and Börner[7] used publications from the Scientometrics Journal in their study, while other researchers have used data retrieved from SCOPUS or Web of Knowledge. Even though data from these sources are arguably the most reliable, it is still necessary to investigate if other sources such as BioRxiv can be suitable alternatives for researchers who may not have access to data from Web of Science or SCOPUS.

As data from BioRxiv is publicly available through Rxivist, applications that aid new researchers or PhD students in finding promising/upcoming research trends can be developed using this data with the mixed indicators model. Thus this study uses these indicators for an exploratory study on Biological Sciences research to determine if:

1. Mixed model indicators can effectively identify emerging research topics from publication data extracted from BioRxiv.

2. Publication data from BioRxiv are suitable for bibliometric analyses.

Subsequent sections give an overview on emerging research, describe the BioRxiv dataset and report on the use of clustering techniques to identify emerging research trends with the mixed indicators model.

BACKGROUND

Rotolo, Hiks and Martin[8] listed core features of an emerging research front as fast growing, novel, coherent and providing beneficial impact and these attributes could be measured through co-word and citation analysis. There is however a consensus that novelty and growth are the main features of an emerging technology,[5] with reduced growth arguably suggesting dying research topics.

To investigate knowledge evolution in scientific disciplines, a top percentage of highly cited articles within a time interval are analysed,[1,2] Small[2] used the top 1 percent highly cited papers in his analysis, but this percentage could be adjusting for varying results.

Quantitative investigation of knowledge evolution by bibliometric techniques are sometimes backed up by qualitative analysis based on interviews of experts in a discipline,[1] using experts to select keywords used in co-word analysis[4] or using experts to identify promising research fronts.

Even though the majority of researches have used techniques such as bibliometric coupling and co-citation, Smalheiser[3] demonstrated that text mining approaches can be used for these analyses. The study[3] queried a database and manually inspected the resulting textual results for possible trends.

Data sources such as Web of Knowledge that are suitable for the investigation in emerging or dying scientific trends are not readily available for public use, as such it becomes necessary to to investigate other reliable data sources. BioRxiv, introduced in 2013 is a preprint database that holds preprints in the life sciences scientific discipline. Researchers usually publish preprints on BioRxiv before they are submitted as manuscripts to journals for peer review and life sciences researchers are increasingly embracing the use of preprint servers such as BioRxiv.[9,10] Thus it is a good source for early detection of emerging topics because it removes the time taken for peer review, publication and ISI indexing processes. Also, it has been shown that up to two thirds of preprints submitted to BioRxiv are subsequently published in peer reviewed journals and there is a correlation between the quality of journal the preprint is later published in and the number of preprint downloads.[10] However, as this data source is relatively new, it is necessary to investigate its suitability for bibliometric studies. Thus, this research investigates ways to extract the mixed indicator metrics[7] for data collected from BioRxiv to determine the extent to which this data is suitable of identifying emerging research trends.

METHODS

Data

Rxivist indexes all submissions to BioRxiv, includes additional altmetrics and makes this data publicly available for download in a structured format,[9] thus it is possible for researchers without access to ISI databases to carry out experiments on emerging research fronts. Complete data from Rxivist was downloaded and then metadata fields; preprint title, preprint abstract and preprint year were extracted and used for the analyses in this study.

Mixed Indicators

This research is inspired from the mixed indicator model approach;[7] however as the Rxivist dataset does not include reference information, the indicators used in the mixed model approach[7] are adapted to:

1. Keyword bursts: Percentage increase in the number of keyword mentions across years.
2. New authors: Increase in authors who use keywords.
3. Interdisciplinarity: Guo, Weingart and Börner[7] have shown that emerging research fronts usually contain references from multiple disciplines, as this study is based on keyword analysis, the extent to which a paper is interdisciplinary is measured by the number of keywords that co-occur with a particular keyword in preprints.

Keywords are integral for the mixed indicators above but the Rxivist dataset does not include keyword information, thus it is necessary to efficiently extract life science keywords from the title and abstracts of bioRxiv preprints.

Keyword Extraction

Rapid Automatic Keyword Extraction (RAKE)[11] is a keyword extraction algorithm that identifies candidates keywords by extracting sequence of words that are delimited by predefined stop words in a text document. The candidate keywords are then scored based on:

- Frequency: The number of times the candidate keyword appeared in the text document.
- Degree: The number of words the candidate key word is made of.
- Degree frequency ratio: The ratio of degree to frequency.

This study uses the degree frequency ratio to extract keywords whose score is greater than one, favouring words with high degree and also excluding candidate keywords that appear only once. The RAKE algorithm does not take the Part of Speech (POS) tag into account when extracting keywords, thus adjectives and verbs which are not likely to be keywords in journal publications may be part of the key phrases. To limit this, the RAKE extracted key phrases which are not part of the noun phrases identified by regular expression matching of POS tags are excluded. An example is shown in Figure 1.

The extracted RAKE keywords in Figure 1 highlights its limitation where phrases like "extensively applied" and "usually genes" appeared as candidate keywords. This can be avoided by introducing a set of stop words that will exclude certain key phrases from or the noun phrase regular expression matching technique that is used in this study. The following patterns were used to match a sequence of POS tags as candidate keywords:

- A sequence of zero or more adjectives followed by a noun.
- A sequence of proper nouns.

The extracted keywords from a document can be used as an identifier of the scientific topics the document belongs to. As keywords are determined dynamically, it increases the possibility of having irrelevant key phrases (noise), however it also makes it possible to capture new trends as opposed to using a predefined keyword list. Ultimately, it will be beneficial for preprint databases to also include keywords in the metadata of preprints for accurate data analyses.

Each extracted keyword as a scientific trend and its mixed indicator score for the years 2013 to 2018 are then used in an exploratory cluster analysis to identify research trends. Table 1 shows the increase in number of authors (author burst), increase number of mentions (keyword bursts) and unique keyword mentions (inter disclipinarity) for the key phrase “genome evolution” that is in the abstract in Figure 1 between 2017 and 2018.

Clustering

Supervised and unsupervised machine learning are techniques used in data analysis. Supervised learning, regarded as classification attempts to assign instances into predefined labels while unsupervised learning, also regarded as clustering enables the grouping of identical instances in a dataset without any predefined label whist ensuring that instances grouped together are more similar that instances in different groups. Clustering is sometimes considered to be more difficult than classification.[12]

Clustering techniques are broadly divided into partitioned or hierarchical algorithms. Hierarchical clustering algorithms such as Chameleon[13] and the Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)[14] and others work by merging or splitting clusters to form a new set of clusters for the subsequent iteration. Hierarchical clustering algorithms that form new clusters by merging are known as agglomerative clustering algorithms while those that form new clusters by splitting are known as divisive. In agglomerative clustering, each instance in the dataset start as singleton clusters and then clusters in iteratively merged until a stopping condition is met. Divisive hierarchical clustering algorithms start as one cluster and then the clusters are iteratively split until a stopping condition is met.

Partitional clustering algorithms such as K Means[15,16] and CLARANS (Clustering Large Applications based on Randomized Search)[17] assign an instance to a predefined cluster which usually the cluster that is the most similar to that instance, where similarity is measured by criterion such as Euclidean distance.[12] These clustering algorithms have been used in exploratory web metrics research[18] and some of these clustering algorithms are implemented in the sckit-learn[19] machine learning python package.

Clustering is useful for exploratory analyses in order to identify trends that may exist in data. A cluster analysis of the mixed indicator dataset in this study may show emerging or dying research trends for the respective topics.

KMeans

The KMeans clustering algorithm is arguably one of the best-known[20] and its simplicity makes it widely used. Given a dataset to be grouped into K clusters, the KMeans algorithm identifies clusters by:

1. **Initialization:** Initial centroids for the K clusters are identified. The centroids can be determine by randomly choosing n instances in the dataset as centroids (Forgy), or assigning instances to a random cluster and then computing the centroids as the average of instances in each cluster (Random Partition).[21]

2. **Assignment:** Instances are iteratively introduced to be assigned to a cluster. An instance is assigned to the cluster with the least distance between the cluster centroid and that instance. The least squared error (Euclidean distance) is a common metric that is used to compare instances with cluster centroids.

3. **Update**: After an instance is assigned to a new cluster, the centroid of that cluster is updated as the average of all instances belonging to that cluster.

Variations to the KMeans exist,[22–25] however, the majority are based on the initialization, assignment and update phases.

Affinity propagation

An important requirement for the KMeans algorithm is determining the initial number of clusters and their centroids as these initial k clusters and their centroids greatly influence the resulting clusters identified by the KMeans algorithm.[22] The affinity propagation clustering algorithm[26] do no need a predefined number of clusters so this algorithm may be used to determine the number of clusters for the KMeans algorithm or can be used as a standalone clustering algorithm for exploratory analyses of a dataset.

In the Affinity Propagation clustering algorithm, all instances are possible cluster centroids (exemplars); and the similarity S(i, j) in matrix S shows the extent to which an instance S_j is a suitable centroid for instance S_i. $S_{j,j}$ which is the likelihood that an instance is a suitable centroid for itself is used to control the number of cluster. For exploratory analyses where nothing is known about the dataset, these initial values ($S_{j,j}$) could be set to a common value; the median of instance similarities.[26]

Two matrices, availability and responsibility are used to determine the final cluster centroids. The availability matrix is initialized to zero while the responsibility matrix is initialized to the similarity matrix S. These availability and responsibility matrices are updated in every iteration of the affinity propagation algorithm as described in previous research.[26] The responsibility information r(i, k) denotes the extent to which an instance k should be a cluster centroid for instance i compared to other possible cluster centroids, while the availability information a(i, k) denotes the suitability of instance k to be a cluster centroid of instance i, given the other instances instance k is a cluster centroid for. The final cluster centroid are those instances with positive availability and responsibility values.

The affinity propagation clustering algorithm is faster and more accurate that other clustering algorithms such as the KMeans.[26]

Abstract extracted from[27]
No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. Phylostratigraphy is a computational framework for dating the emergence of sequences (usually genes) in a phylogeny. It has been extensively applied to make inferences on patterns of genome evolution, including patterns of disease gene evolution, ontogeny and de novo gene origination. Phylostratigraphy typically relies on BLAST searches along a species tree, but new simulation studies have raised concerns about the ability of BLAST to detect remote homologues and its impact on phylostratigraphic inferences. These simulations called into question some of our previously published work on patterns of gene emergence and evolution inferred from phylostratigraphy.
Extracted RAKE keywords (Degree frequency ration greater than one)
BLAST - BLAST searches along - computational framework - de novo gene origination - detect remote homologues - disease gene evolution - emergence - evolution - evolution inferred - extensively applied - gene emergence - genome evolution - including patterns - make inferences - new simulation studies – patterns - phylostratigraphic bias impacting inferences - phylostratigraphic inferences – phylostratigraphy - phylostratigraphy typically relies - previously published work - raised concerns - simulations called - species tree - usually genes
Regular expression matched noun phrases
Ability - BLAST - computational framework – disease - emergence - evidence - evolution – gene - genome evolution – impact - new simulation – ontogeny – origination – phylogeny - phylostratigraphic bias – phylostratigraphy – question – tree - work
Final extracted keywords (intersection of regular expression matched noun phrases and RAKE keywords)
BLAST - computational framework – emergence – evolution - genome evolution - phylostratigraphy

Figure 1: An example of keywords extracted from an abstract using the RAKE algorithm and regular expression noun phrase matching.

RESULTS AND DISCUSSION

Result of clustering algorithms may group topics into clusters, to determine the quality of the clustering solution, intra and inter cluster similarity is used. Ideally, the inter cluster similarity should be high while intra cluster similarity should be low. This evaluation metric has been previously used in researches.[18] Table 2 shows the quality of the clustering solutions identified with the KMeans and Affinity Propagation algorithms. As the affinity propagation algorithm dynamically identifies the number of clusters, the number of clusters identified by the affinity propagation algorithm is used as input for the KMeans algorithm.

Silhouette,[28] Davies Bouldin[29] and Carlinski-Harabasz[30] (metrics) are used to determine how well the clustering algorithms grouped keywords based on their mixed indicators.

This silhouette score is determined by the average distance between an instance and other instances in its assigned cluster (within cluster similarity) minus the average distance between an instance and other instances in the cluster most similar to its assigned cluster (neighbouring cluster similarity) divided by the maximum between the within cluster similarity and neighbouring cluster similarity.[28] The final score bounded between -1 and 1 is then the average silhouette score for each instance in a dataset. Lower silhouette scores indicate worse clustering solutions, which reflects the extent to which an instance in a clustering solution is assigned to its most appropriate cluster.

The Carlinski-Harabasz (CH) score also regarded as the variance ratio criterion[30] is determined by the ratio of the sum of differences between instances and their corresponding cluster centroids (Within Cluster Similarity) to the variance between cluster centroids and the centre of the dataset (average of all instances). The Within Cluster Similarity (WCS) increases as the number of clusters increases, while the Between Cluster Similarity (BCS) reduces as the number of clusters increases. High WCS suggests compact clustering solutions where instances are close to their cluster centroids

Table 1: **Extracted mixed indicator values for key phrase "genome evolution" in the Rxivist dataset.**

Keyphrase	Author Burst	Interdisciplinarity	Keyword Burst
Genome Evolution	1.71	4	-0.65

Table 2: **Quality of clustering solutions for the mixed indicator dataset clustered with the KMeans and Affinity Propagation (AP) algorithms.**

	Silhouette Score	Davies-Bouldin Score	Carlinski-Harabasz Score
KMeans	0.254	0.791	5420.44
AP	0.241	0.786	5376.37

and high BCS increased dispersion between cluster centroids. Thus, a higher CH score is a better clustering solution as it indicates that clusters in the clustering solution are compact and the centroids are dissimilar.

The Davies-Bouldin (DB) score evaluates the quality of a clustering solution as the ratio of the average within cluster distances to the average distances between the cluster centroids in clustering solution, where the distance is measured by a metric such as the Euclidean distance. As a good clustering solution should have low within cluster distances and high between cluster distances, lower Davies-Bouldin scores suggests better a clustering solution.

The initial similarity matrix in the Affinity Propagation clustering algorithm controls the number of clusters that will be identified.[26] In tests, larger values of the initial similarity resulted in a singleton clustering solution, where clusters comprised of only one instance and as a consequence, better silhouette scores. However this may not be regarded as a good clustering solution as instances are not grouped. When the values of the initial similarity matrix increases, the number of clusters (groups) identified by the Affinity Propagation Algorithm also increases. As shown in Figure 2, an appropriate initial similarity value for the dataset in this study is in the neighbourhood of -300, where the quality of the clustering solution as determined by the silhouette evaluation metric is comparable to a clustering solution of singleton clusters. Thus the results reported in Table 2 is the quality when the similarities are initialized to -300 for the Affinity Propagation Algorithm and k = 46 (the number of clusters for the Affinity Propagation Algorithm when the similarity is initialized to -300) for the KMeans algorithm.

The visualization of the keyword burst vs. author burst between 2017 and 2018 in Figure 3 showed some visible separation of groups that formed unique clusters with the KMeans clustering algorithm. The analysis is limited to submissions between 2017 and 2018 because the number of bioRxiv submissions in earlier years are too few for meaningful analyses.

As there are three indicators; keyword burst, author burst and interdisciplinary, a Principal Component Analysis (PCA) was performed to enable plotting a 2D graph. Figure 3 is a plot of the first and second PCA components.

On investigation of emerging or dying research fronts, outliners play an important role in forming distinct cluster areas. The top area of the graph in Figure 3 contained keywords with negative change in author and keyword bursts between 2017 and 2018 (reduced interest), the bottom area contained keywords with positive change in author and keyword bursts between 2017 and 2018 (increased interest), while the right of the graph showed keywords with increased interdisciplinarity.

Keywords that appeared in the interdisciplinary cluster area were generic key phrases such as

"*Study, method, analysis, model and disease.*", thus, this may simply be a consequence of the increased number of publications between 2017 and 2018. It may also be necessary to identify a more appropriate approach to keyword extraction compared to RAKE[11] that is used in this research which will exclude

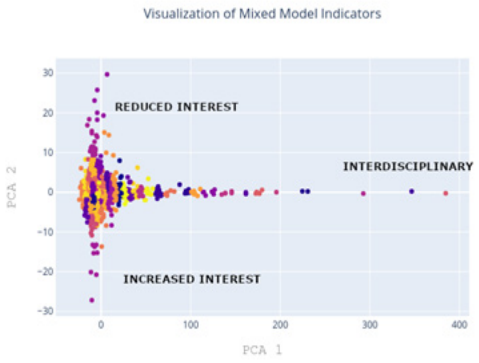
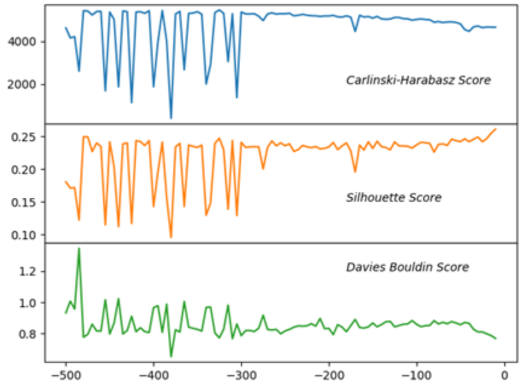
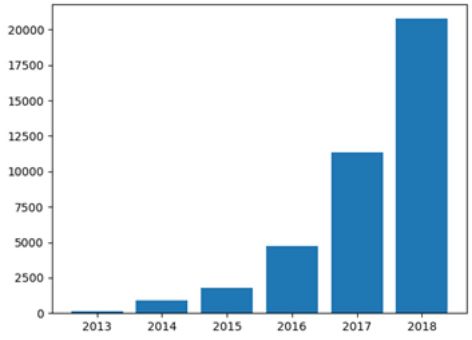


Figure 3: Visualization of KMeans identified clusters of mixed model indicators.



s. ilarities

Figure 2: Quality of clusters identified by the Affinity Propagation Algorithm for the mixed indicator dataset with varying initial similarities



imissions from 2013 – 2018

Figure 4: BioRxiv preprint submissions from 2013 to 2018.

these generic terms. However, researches such as this study will benefit from bioRxiv including keywords as part of its metadata.

Increased keyword and author bursts cluster area contained keywords such as *"breast, prostate cancer, ovarian cancer"* suggesting increased interest in cancer studies.

The reduced interest cluster area contained keywords such as *"Brazil, European ancestry, cardiovascular disease".* On inspection of the abstracts, the only identified pattern was the relationship between Zika virus and Brazil in 2017. Interestingly, even though "zika" virus appeared in subsequent years, it did not co-occur with *"brazil",* hence the reduced number of publications with "brazil" as a keyword.

The number of preprints submitted to BioRxiv increase significantly every year with the number of submissions in 2018 more than the submissions from 2013 to 2017. This may not necessarily suggest increase in research activities in the life sciences discipline but highlights the increased awareness of the database among life science researchers. The chart in Figure 4 suggests that the Rxivist database is arguably not yet mature enough for reliable data analyses, so the results from this study should be taken with caution. However, following the increased submission trends, Rxivist database along with the methods described in this research will be valuable in identifying interesting trends for emerging and/or dying research fronts in the life sciences discipline.

CONCLUSION

It is the consensus that investigating emergence or death of research fronts is important and state of the art approaches use bibliometric techniques such as co-word and citation analysis of of data extracted from ISI databases. As data from ISI databases are not publicly available, it becomes necessary to investigate the suitability of alternative sources, for example BioRxiv which stores preprints from life sciences scientific discipline.

It is possible to freely download metadata (title, abstract, author) and other metric such as download counts and twitter analytics of BioRxiv preprints from Rxivist. As preprints are published to BioRxiv before submitted to journals for peer review, accepted, published and then indexed in ISI databases, Rxivist dataset not only has the advantage if being freely available for researchers, it can also identify the emergence of topics earlier than data from ISI databases.

Rxivist dataset does not contain keywords in its metadata, hence it is necessary to extract keywords from the abstract and title of preprint submissions. To automate the keyword extraction process for co-word analysis, candidate keywords in this study are an intersection of keywords identified by noun phrase regular expression matching with key phrases identified by the RAKE[11] keyword extraction algorithm.

Mixed indicator models have been shown to be able to identify emerging research trends.[7] Keywords extracted from the Rxivist dataset were transformed into mixed indicator models (keyword burst, author burst and interdisciplinary) and further clustered using the KMeans and Affinity Propagation algorithms to identify any possible existing trends.

Outliners in the mixed indicator model dataset which also formed clusters are candidates for further inspection to determine emergence or death of research topics. A cluster (outliner) in the dataset that suggested emerging fronts (increased author burst), was in cancer research. However as the BioRxiv database is still relatively new, with the number of preprint submissions approximately doubling yearly, author burst could simply indicate that more authors in a research topic are embracing the use of the preprint database, which Abdil and Blekhman[10] has shown to be on the rise in the life sciences discipline. The current lifetime of BioRxiv is arguably not sufficient to investigate the novelty or growth of research topics, but as the BioRxiv dataset becomes more mature in the coming years, the methods described in this research will be valuable for the investigation of research topic emergence or death.

This study has demonstrated the use of text mining approaches to investigate the suitability of Rxivist dataset for identifying emerging or dying research trends, however other metrics such as download count that are also part to the Rxivist dataset may be incorporated into the analysis. It may also be beneficial for keywords to be part of the metadata in the preprint submissions to BioRxiv, which will improve the accuracy of co-word analysis as it removes the requirement of automatic keyword extraction.

ACKNOWLEDGEMENT

The author would like to thank the reviewers for their useful comments

CONFLICT OF INTEREST

The author declares no conflict of interest.

ABBREVIATIONS

BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies; **WCS:** Within Cluster Similarity; **BCS:** Between Cluster Similarity; **RAKE:** Rapid Automatic Keyword Extraction.

REFERENCES

1. Upham SP, Small H. Emerging research fronts in science and technology: Patterns of new knowledge development. Scientometrics. 2010;83(1):15-38.

2. Small H. Tracking and predicting growth areas in science. Scientometrics. 2006;68(3):595-610.
3. Smalheiser NR. Predicting emerging technologies with the aid of text-based data mining: The micro approach. Technovation. 2001;21(10):689-93.
4. Lee WH. How to identify emerging research fields using scientometrics: An example in the field of Information Security. Scientometrics. 2008;76(3):503-25.
5. Small H, Boyack KW, Klavans R. Identifying emerging topics in science and technology. Res Policy. 2014;43(8):1450-67.
6. Huang MH, Chang CP. A comparative study on detecting research fronts in the organic light-emitting diode (OLED) field using bibliographic coupling and co-citation. Scientometrics. 2015;102(3):2041-57.
7. Guo H, Weingart S, Börner K. Mixed-indicators model for identifying emerging research areas. Scientometrics. 2011;89(1):421-35.
8. Rotolo D, Hicks D, Martin BR. What is an emerging technology?. Res Policy. 2015;44(10):1827-43.
9. Abdill RJ, Blekhman R. Rxivist.org: Sorting biology preprints using social media and readership metrics. Plos Biol. 2019;17(5):1-10.
10. Abdill RJ, Blekhman R. Tracking the popularity and outcomes of all bio Rxiv preprints. Elife. 2019;8.
11. Rose S, Engel D, Cramer N, Cowley W. Automatic Keyword Extraction from Individual Documents. In: Text Mining. John Wiley and Sons, Ltd. 2010;1-20.
12. Saxena A, Prasad M, Gupta A, Bharill N, Patel OP, Tiwari A, *et al.* A review of clustering techniques and developments. Neurocomputing. 2017;267:664-81.
13. Karypis G, Han EH, Kumar V. Chameleon: Hierarchical clustering using dynamic modelling. Computer. 1999;32(8):68-75.
14. Zhang T, Ramakrishnan R, Livny M. BIRCH. In: Proceedings of the 1996 ACM SIGMOD international conference on Management of data-SIGMOD '96. New York, New York, USA: ACM Press. 1996;103-14.
15. Arthur D, Vassilvitskii S. K-means++: The Advantages of Careful Seeding, SODA'07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, New Orleans, Louisiana, 2007." URL https://theory. stanford. edu/~ sergei/papers/kMeansPP-soda. pdf: 1027-1035.
16. Macqueen JB. Some methods of classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967;281-97.
17. Ng RT, Han J. CLARANS: A method for clustering objects for spatial data mining. IEEE Trans Knowl Data Eng. 2002;14(5):1003-16.
18. Kenekayoro P, Buckley K, Thelwall M. Clustering research group website homepages. Scientometrics. 2015;102(3):1-14.
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* Scikit-learn: Machine Learning in {P}ython. J Mach Learn Res. 2011;12:2825-30.
20. Lam D, Wunsch DC. Clustering. Acad Press Libr Signal Process. 2014;1:1115-49.
21. Hamerly G, Elkan C. Alternatives to the K-means Algorithm That Find Better Clusterings. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management. New York, NY, USA: ACM. 2002;600-7. (CIKM '02).
22. Erisoglu M, Calis N, Sakallioglu S. A new algorithm for initial cluster centres in k-means algorithm. Pattern Recognit Lett. 2011;32(14):1701-5.
23. Aharon M, Elad M, Bruckstein A. K-SVD: An algorithm for designing over complete dictionaries for sparse representation. IEEE Trans Signal Process. 2006;54(11):4311-22.
24. Chinrungrueng C, Sequin CH. Optimal Adaptive K-Means Algorithm with Dynamic Adjustment of Learning Rate. IEEE Trans Neural Networks. 1995;6(1):157-69.
25. Lu Y, Lu S, Fotouhi F, Deng Y, Brown SJ. Incremental genetic K-means algorithm and its application in gene expression data analysis. BMC Bioinformatics. 2004;5.
26. Frey BJ, Dueck D. Clustering by passing messages between data points. Science (80-). 2007;315(5814):972-6.
27. Domazet-Lošo T, Carvunis AR, Albà MM, Šestak MS, Bakarić R, Neme R, *et al.* No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. Bio Rxiv. 2016;060756.
28. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53-65.
29. Davies DL, Bouldin DW. A Cluster Separation Measure. IEEE Trans Pattern Anal Mach Intell. 1979;PAMI-1(2):224-7.
30. Caliński T, Harabasz J. A dendrite method for cluster analysis. Commun Stat. 1974;3(1):1-27.