Long term Impact of a Research Article

ABSTRACT

The attention a scholarly article receives depend on many factors which include popularity of the author, the research area, the technologies used, and how the research will be beneficial to society. Nowadays, increasing number of articles are being shared on various online platforms. The metrics of these platforms can help in evaluating how an article receives online attention. In this paper, we use these metrics to evaluate the long term impact of an article online. Having the knowledge of how long an article would have an impact could help authors, libraries, publishers, research communities, and funding agencies in determining if an article is worthy to spend the resources and funds towards publishing the article. In our study we use Altmetrics dataset and we build various models on it to predict the impact of an article using various online platforms.

Index Terms- Long Term impact, Online Age of articles

INTRODUCTION

The rise of the internet in recent years have had an increased amount of articles being published throughout the internet especially on social media platforms, increase in the number of scholars to discuss and share their research on various social media platforms has inferred that an article's life span in the online world can play a huge role in determining the article's impact worldwide. Having the knowledge of how long an article would have an impact could help authors, libraries, publishers, research communities, and funding agencies in determining if an article is worthy to spend the resources and funds towards publishing the article. Majority of the articles published would sometimes have a very low lifespan and the effort put in the research article would have less recognition. The prediction of the lifespan of any article would help in avoiding that issue and provide insight as how to properly move ahead with the publication of the article. Predicting the impact of the paper will definitely help the authors in writing a good research paper and their work can be witnessed by the people within that time-span without wasting the valuable resources that were spent before/during/after publishing the research work.

PREVIOUS WORK

Davis and Cochran [1] explained that one of the metrics to measure the impact of an article is the Cited Half-Life. It is observed that mean cited half-life of the scholarly literature is 6.5 years and growing at a rate of 0.13 years per annum. Garfield [2] showed that each discipline will have their own protocol regarding citation. Lariviere et al. [3] observed that the age of cited material has risen continuously since the mid-1960s. Article-level metrics (altmetrics) have been proposed as a complement to scholarly metrics such as citations, the Impact Factor (IF), the immediacy index, and the h-index. There is an increase in sharing of articles on social media platforms like Facebook, Twitter and reference managers like Mendeley, CiteULike and information providing websites like Wikipedia, blogs, news. Altmetric provides a dataset that has all the metrics of scholarly articles on the mentioned online platforms. Lariviere et al. [3] observed that the typical citation life cycle of an article starts with a rapid increase during their initial years on the scientific scenario, followed by a peak, and then a slow but steady fall into oblivion. Barnett and Fink [4] found that the invention of the internet has had an impact on scholarly articles and has increased the length of average life of academic citations by 6 to 8 months, fueling the statement that the internet has had an impact on articles and its impact will continue to increase in the coming years. Martin et.al [5] used cited half life JCR for their work and The article states the recent increase in older articles being cited is due to technology. Tonia et al. [6] have done several studies investigating the relationship between social media metrics and citations.Cunningham and Bocock [7] found that the quantitative measurements such as median citation age and publication half life for a particular and characteristics of a author play a important role in determining the documents rate of aging. Jones et al. [8] says that age-creativity relationship varies frequently over time than across fields and also they analysed the independent and strong association of age dynamic within fields and both the training patterns and prevalence of theoretical contributions.

Old papers are starting to get more and more citations. There might be many reasons for old papers to receive more and more attention after many years of their publication. Šember et al. [9] explored the reason for the regain of attention of old papers is that the technologies are evolved and the authors cite the old papers in order to show the evolution in the technology. Another reason for citation of old papers is their archival value. Oppenheim and Renn. [10] observed that 40 percent of old paper citation is because of the their historical value. The ageing related aspect of information was taken into consideration by Zhang and Glanzel [11] in their study. Synchronous (alias retrospective) approach was followed in which a paper is taken and other papers are back tracked on the basis of citations the particular paper has. It was found that the share of more recent references was distinctly lower in 2014 than that in 1992, which holds for all aggregation levels. Further their study also showed that the "Price Index" is very important measure at the paper level for evaluating the differences between the documents published in the same journals and (sub-)fields. Egghe et al. [12], studied the rate of obsolescence for articles. They observed that the rate of obsolescence vary and this variation was calculated as utilization functions in their study which derived that, in the synchronous case, the larger the increase in production, the larger the obsolescence. In the diachronous (prospective) case the opposite relation holds: the larger the increase in production the smaller the obsolescence rate. Glänzel et al. [13] introduced the term "Prospective Price Index" that expresses the share of citations received in an initial period after publication in those received in a larger citation window and it may serve as a proxy for literature ageing in the mirror of citation processes. Altmetric is a system that tracks the attention that research outputs such as scholarly articles and datasets receive online. Altmetric.com provides the data which shows the online impact the scholarly articles receive. In a study by Thelwall and Nevill [14] the features in the Altmetric.com dataset were used to predict the importance/attention a paper can get in terms of future citation counts. They found that Mendeley reader counts are constant indicators of future citation count an article will receive. Stern [15] found that the ranking of an article can be determined by initial citations in early stages of an article and can also determine the future citations of an article. Brookes [16] has stated that "the faster the rate of growth, the less is the scatter and the more rapid the obsolescence", that those articles which were most productive initially would have short lives and articles which had less productivity would have high active life in the later stages. Stern and Wallace [17] found that the high journal median citation ages were always associated with journals that were unproductive in terms of numbers of references to those journals in the database. Sikdar et al. [18] found a peer review system in which network-reviewer interaction network will be able to predict the long term impact of the paper with the help of basic network features such as degree,clustering coefficient,centrality. Sun et al. [19] proposed a vector for measuring obsolescence of scientific articles O = (Gs,A-,n) where G is an index revealing the history of citations ,A- is a parameter for uncovering citation leaping and age n is an adjusting parameter. Nansen [20] analyses how the Impact Factor (IF) across research categories is correlated with age and number of citations. Their analysis indicates that IF's are positively correlated with amount of published studies and negatively correlated with average age of citations.

Larivière et al. [21] collected data from medical and natural sciences and engineering fields and studied that to have major findings stating that during world wars there was a very slow publication of scientific production, the importance of articles aged 1 year had their importance decreased, 2 years aged showed no activity and 0 years did not increase at all. Luwel and Moed [22] in their study analyse the effect of publication delay on the age of the article. These delays are more in mathematical and technical sciences when compared to other fields. The results of their study shows that the cited half life may reduce by a factor of 2 if the publication delays decrease to a great extent. Davis [23] in his study found that There is a general increase in the cited half-life of the scholarly literature, some journals experienced decrease in cited half lives and cited half-life is related to the total citations, and as a journal attracts more citations much of them are citing old literature. The increase in cited half life is due to industry wide transformation, mass digitization of journals, full text indexing and better search engines.This might also reflect major structural shifts in the way science is funded and the way scientists are rewarded. Hajra and Sen [24] studied few citation networks to find out the age dependence of the attachment probability. According to their results they conclude that majority of papers have a fair chance of getting cited within ten years of publication, after which fewer survive the 'test of time'. According to their study, Egghe and Rousseau [25] say that Once we know the growth distribution,we could correct an observed citation distribution. They show that the main factor is the difference between average and global aging that depicts the influence of growth on aging over a complete period as a whole. Simkin and Rowchowdhury [26] improvised their previous work which only considered that the citation distribution is generally from the list of references used in other papers. In the improvised research they show their findings by considering citations o f paper from 'recent' papers. So according to the novice model

presented, it can be seen that unlike stochastic model to study the literature ageing, this model considers rate of citation as proportional to number of citations a paper receives in previous year. Egghe[27] in his study noticed the flaws of Wallace's[28] work and improvised the work. Wallace's work showed relation between journal's median citation age and its number of articles. Egghe replaced the median with mean in his study. Yao et al. [29] studied the Community Question Answering (CQA) to see their long term impact as soon as they are posted. They proposed various algorithms. For this purpose they model three important aspects: non linearity, question/answer coupling and dynamics. They conduct experiments with their algorithms on two real CQA datasets to evaluate effectiveness and efficiency of their algorithms. Stegehuis et al. [30] in their study proposed Quantile based regression models to predict the future of citations. They used bi-variable (impact factor and early citation counts) models. Their models performed the best when used both of the variables together instead of using them separately.

Wang et al.[31] in their study, made a mechanistic model for the citation dynamics of papers from various journals and disciplines. They made a single curve for various journals and disciplines which showed that all papers tend to follow a same universal temporal pattern. Redner [32] in his work suggested that Since almost all papers are gradually forgotten, the probability that a given paper is cited should decrease in time with a relatively short memory. Datta et al. [33] in their study, determine the half life of software engineering research topics taking into account over 19,000 papers from software engineering publication venues during 1975-2010. They used natural language processing to identify and associate topic to a paper. The results show the some research topics have a half life of nearly 15 years. Singh et al. [34] analyzed the effects of various types of citations at different points in the lifestyle of an article. They compared effects of "influential" and "non-influential" authors citing a work early in the life cycle. They also compared effects of early self-citations, co-author citations, and more citations by more distant authors. They found a *negative* correlation between early citations by high-impact authors and long-term citation count. They suggest that the higher-impact authors may "steal the glory" of the original article. They also created a model that confirmed their idea of early influential citers having a negative effect on the citation counts of an article. Matricciani [35] proposed entropy H of age T of citation in literature. He derived parameter S=exp(H) that measures obsolescence of literature and discussed the mean residual life M(T), the expected life E(T) of reference of age T which may be used to define age of historical/old papers.

INNOVATION

Our study is the first to predict how long a given article may remain relevant by using altmetric data and predicting the lifespan of any article through the means of online.

DATA

A random sample of data was collected from altmetrics dataset which contained over 2,100,000 records. We have used java code for extraction of data from the json files and stored in MySQL database. Then we extracted the data into a csv to perform model building tasks. The features included in our research is provided in Table 1. The features in Table 2 were used to create two other features which are described in Table 3.

Features	Description
Altmetric ID	Altmetric ID of research article
Mendeley Count	The amount of reader counts on Mendeley platform.
CiteuLike count	The amount of reader counts on CiteuLike platform
Connotea count	The amount of reader counts on Connotea platform
Blogs count	The amount of reader counts on Blogs
News Count	The amount of reader counts on News
Twitter Count	The amount of reader counts on Twitter platform
Wikipedia Count	The amount of reader counts on Wikipedia platform
Facebook Count	The amount of reader counts on Facebook platform
Google + count	The amount of reader counts on Google+ platform
Policy count	The amount of reader counts on Policy

Table 1. Features and their Description

Features	Description
Publish Date	The published date of the article
Blog posted Last Date	The last posted date of the article on blogs.
Twitter posted Last Date	The last posted date of the article on Twitter

Google + posted Last dates	The last posted date of the article on Google+
Wikipedia posted Last dates	The last posted date of the article on Wikipedia
news posted last dates	The last posted date of the article on news
policy posted last date	The last posted date of the article on policy
QnA posted last dates	The last posted date of the article on QnA
Reddit Post Last Dates	The last posted date of the article on Reddit
Facebook Posted Last Dates	The last posted date of the article on Facebook

Table 2. Features used for Feature Creation

Features	Description
Last Posted date	The last posted date of article online
Online Age	Years the article was active online

Table 3. Features Created

All the last posted dates described in Table 2 were used to generate a new feature called 'Last posted date', this feature gives information about the date the article was last posted online. Using this new feature and the publish date of the article, we created another new feature called 'Online age of the article' which is the number of years the article was active on online social media platforms from its publication date. The main features of our research were the features in Table 1 and Table 3.

DATA CLEANING

The data initially was around 2,100,000 articles with respective to many features When dealing with the big data there is a high chance of duplication of records. Altmetric ID is the unique feature that differentiates one article from another article. With the help of ID the problem of duplication of records was solved. It was noted that some of the records did not have any posted dates on any of the social media platforms and some of the articles had invalid publication dates. We were only interested in articles from the year 1920 to 2009 because not many articles before 1920 contributed to the dataset and articles after 2009 were hard to include since the online age for the articles in this timespan would be very low and it would be difficult to properly build a model with such a complex data. After cleaning of the data by the above criteria we had a dataset of 815,438 records.

When we created a new feature by calculating the difference between the last posted date of an article on social media and published date of the paper for the articles in our dataset, we found that we had some negative values. Inorder to check why we had negative values we evaluated our method of extraction thoroughly and found no error in it so we checked the data itself provided by the altmetrics here we found the error Incorrect data may lead to the bad decision we removed such records,giving us a final dataset of 673,530 records. With this dataset we performed classification models on this data. we observed a low f1-score for the models we considered. this is why because features ranged widely throughout the years. the period of 19th century(1920-1999) has fewer records when we compared to the period of 20th century(2000-2009). hence this is a outlier problem the solution we found was to make clusters of years with respect to the features. With this ,we started proceeding with the further process.

METHODOLOGY

The main idea of our research was to use regression models to predict the lifespan of each article. but analysing that the research was a time series issue, we decided to perform classification but we noticed that the results we got were very poor since the data varied a lot from different years and hence we decided to form clusters of different years based on the features we had extracted, then based on the clusters which are formed build appropriate classification models. Firstly, we calculate the difference between the last online mentioned data and the published date of the article i.e the online age of said article. Secondly, we calculate the median of these online ages. The calculated median can be used to compare the other online ages and finally we make classification models in order to see if the predicted values is greater or equal to median of the online age. In detail, data should be clustered in such a way that there should be minimal noise in the clusters we formed.

Defining the number of clusters was very troublesome. For this problem we took random number of clusters after visualizing Total Online counts vs years. randomly we took the number of clusters and applied the models and checked the accuracy. if we found it was not good then increased the number of clusters by value of one. We used k means clustering to form different clusters based on the sum of the counts of different platforms of each year. We ended up on choosing 5 as the k value since it better suited the ideology behind our research. We conducted k means algorithms many times on our data to come up with an appropriate value of k and to check all the clusters

are in the same position since k means starts by taking the random point and performs the clusters based on the points it took so, for every run it will give different results. Figure 1 shows the clustering of the articles with k as 5.

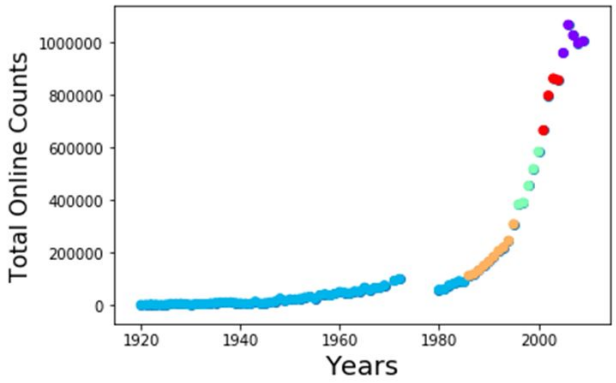

Fig 1. Clusters

Each cluster and their publication year that was analyzed are provided in Table 4.

Cluster Number	Years abstracted from
Cluster 1	1920-1985
Cluster 2	1986-1995
Cluster 3	1996-2000
Cluster 4	2001-2004
Cluster 5	2005-2009

Table 4 - Clusters formed and their publication date

For each cluster thus formed we computed the median of the online age of the article. We decided to use median as the main criteria since it represents the center of the data and is susceptible to outliers. Our next step is to build classification models for each of the five cluster and see if the model predicts whether the article will have high lifespan online or not based on the median.

CLUSTER 1:

We start the building of models from the first cluster which included articles from the years 1920-1985.This cluster had 195,992 records and surprisingly Mendeley counts were contributing much to the data, the articles might have been posted on the said platform during the 2000's, also the policy counts were important to the data. Through some analysis we found that the most important features included in this cluster was the Mendeley and policy counts, but considering all the features we built different classification models and selected the models which had high F1 - score. Table 5 describes the precision, recall and f1-score of the four models which were performed on cluster one.

Model	Precision	Recall	F1-score
Decision Tree Classifier	0.77	0.75	0.75
Bernoulli NB	0.77	0.76	0.76
Random Forest Classifier	0.77	0.76	0.75
Logistic Regression	0.78	0.76	0.76

Table 5. Results from Cluster 1

From the table we can say that almost all the models have similar F-1 scores, but Random Forest and Logistic Regression performed slightly better with an F-1 score of 0.76. We then built models on the second cluster.

CLUSTER 2:

This cluster had 116,512 records and the cluster two contained articles from the year 1986 to 1995. This cluster had a lot of ups and downs since this was the timespan when the internet started to grow. The most important features for this cluster were again Mendeley and policy counts. We then applied the models on the cluster and had results as described in table 6.

Model	Precision	Recall	F1-score
Decision Tree Classifier	0.82	0.79	0.78
Bernoulli NB	0.84	0.78	0.77
Random Forest Classifier	0.82	0.79	0.78
Logistic Regression	0.82	0.75	0.73

Table 6. Results from Cluster 2

This cluster had The models built on cluster two also had somewhat similar results ranging mostly from 0.73 to 0.78. Decision Tree and Random Forest had the highest F-1 score of 0.78 outperforming Bernoulli by 0.01. Logistic Regression performed had a low F-1 score among all the models with 0.73 F-1 score. Noting that Random Forest has performed well in both the clusters. With the analysis of the results we moved on to the third cluster.

CLUSTER 3:

Then we build the models for cluster 3 which ranges from 1996-2000. This cluster had 98047 records and the most important features we found for this cluster were again mendeley counts, twitter counts and policy counts. From this trend we can deduce that Social media started to rise in this period and most of the articles started to appear on social media in this time duration.

Model	Precision	Recall	F1-score
Decision Tree Classifier	0.84	0.82	0.82
Bernoulli NB	0.85	0.82	0.82
Random Forest Classifier	0.84	0.82	0.82
Logistic Regression	0.83	0.77	0.77

Table 7. Results from Cluster 3

As we can notice from table 7, the models built on cluster three performed very similar with an F-1 score of around 0.82 for the models except Logistic Regression. Logistic Regression performed with a very low F-1 score of 0.77 when compared to the other models. Random Forest has been consistently performing very well since the first cluster.

CLUSTER 4:

We then moved on to cluster four which had articles from the year 2001-2004. This cluster had 102,665 records and for this cluster we observed that Mendeley counts, twitter counts and policy counts are again the important features, since we know that the trend of social media has increased and a lot of research articles were being posted on various social media platforms, since we considered all the features and the values in it were contributing to the data it is obvious that the models that we choose for classification will definitely have a better precision and recall.

Model	Precision	Recall	F1-score
Decision Tree Classifier	0.83	0.82	0.82
Bernoulli NB	0.85	0.83	0.83
Random Forest Classifier	0.83	0.82	0.82
Logistic Regression	0.82	0.77	0.77

Table 8 - Results from Cluster 4

The models built on the fourth cluster can be seen from table 8. Similar to the previous cluster the three models had similar results but this time Bernoulli Naive Bayes outperformed other classifiers by a small range.. The precision ,recall and f1 score is similar to the decision tree classifier and random forest classifier it leads to the assumption that random forest classifier is ensemble trees that derived from the decision tree concept. Logistic Regression again underperformed when compared to the other classifiers.

CLUSTER 5:

The final cluster consisted of articles from the years 2005 to 2009 and had 160,314 records, all the features contributed to the dataset for this cluster. But, the most important feature was Mendeley again. We applied various models again and displayed the results for the top models in table 9.

Model	Precision	Recall	F1-score
Decision Tree Classifier	0.77	0.76	0.76
Bernoulli NB	0.79	0.77	0.77
Random Forest Classifier	0.77	0.76	0.77
Logistic Regression	0.73	0.74	0.73

Table 9 - Results from Cluster 5

The results for this cluster was on an average close to 0.77 F-1 score, this was basically due to a lot of variation in the data, since twitter, facebook and other sites saw an increase in their usage. All the models here had similar results where only Bernoulli had a 0.01 decrease, all the rest of them performed similarly to each other. Random Forest consistently performed better over all the clusters. Decision Tree also performed slightly well over all the clusters.

RESULTS

After building models on all the clusters we took the average of each classifier to analyse which model has performed well overwell . The results can be found in table 10.

Model	Precision	Recall	F1-score
Decision Tree Classifier	0.8	0.78	0.78
Random Forest Classifier	0.806	0.79	0.79
Bernoulli	0.82	0.79	0.788
Logistic Regression	0.79	0.75	0.75

Table 10 - Average results of all the clusters

From analysing the above table we can conclude that Random Forest classifier had performed slightly well when compared to the other classifiers. Logistic Regression has performed poorly over all. We think random forest has performed well because its strength is

feature importance,. it selects the features based on the importance and builds the trees based on these features which are less noisy. Logistic Regression performed well during the initial clusters but then underperformed for the other clusters. These days, enormous number of articles are being shared on different online stages, the measurements of these stages can help in assessing how an article gets online consideration. In this paper, we utilize these measurements to assess the long term impact of an article on the web. Predicting to what extent an article would have an effect could help writers, libraries, distributors, explore groups, and subsidizing organizations in deciding whether an article is qualified enough to spend the assets towards publishing the article. Data cleaning was as an arduous task in our study since there are all types of uncleaned data was present. Predicting the long term impact of the paper is a time series problem since the data varies over time. Thus, we opted for a clusterization technique to evenly distribute the noise for making the performance of the model better. We conducted Grid Search using scikit-learn GridSearchCV which stands for grid search cross validation to tune the hyperparameters. However, most of the models were not improving a lot and hence disregarded this theory.

CONCLUSION

Random Forest Regression has performed effectively well when compared to other models. It chooses features in the view of significance and assembles the trees in light of these features which are less noisy. Decision Tree and Bernoulli Classifiers also performed well and provided results quite similar to Decision Tree. Logistic regression performed well for the beginning clusters then failed to meet expectations for alternate clusters.

We can conclude that we have successfully created a model which predicts the long term impact of a paper on social media. Through Detailed analysis of the results we can conclude that Mendeley counts are the main key in determining the long term impact of an article online. Policy counts also contributed a lot to the data. Random Forest classifier has out-performed when in comparison with the other classifiers.

FUTURE WORK

Our future work includes collecting the data from 2010 to present and apply the model we have chosen previously. We want to do time series analysis on this data, to ensure what results we would obtain and compare those results with the research performed in this paper. We would like to perform time series by making the time series stationary, estimating and eliminating trend, seasonality and forecasting a time series. This model should work on any time period we choose irrespective of the internet usage in that period. Then we can foresee the effect of the paper will enable the writers in composing a decent research paper and their work can be seen by the general population inside that time-traverse without squandering the important assets that were spent previously/amid/subsequent to distributing the exploration work. We would like to do regression analysis collectively as well to predict the exact number of years the research article would have impact online. Finally we want to train our model on a regular basis using fresh data and automate this process as much as we can. In addition to online metric counts such as google plus, citeulike, mendeley, twitter, facebook we want to include the number of citations of each research article and also the h-index of authors to evaluate if our model performs better in comparison to the research done in this paper.

REFERENCES

[1] Philip M. Davis, Angela Cochran, “Cited Half-Life of the Journal Literature”, 2015. http://publishers.org/sites/default/files/uploads/PSP/journalusagehalflife.pdf

[2] Garfield, E. 1975. “The 'obliteration phenomenon' in science - and the advantage of being obliterated!” Essays of an Information Scientist, Vol. 2: 396-398.

[3] V. Larivière, É. Archambault, and Y. Gingras, “Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004),” J. Am. Soc. Inf. Sci. Technol., vol. 59, no. 2, pp. 288–296, 2007.

[4] G. A. Barnett and E. L. Fink, “Impact of the internet and scholar age distribution on academic citation age,” J. Am. Soc. Inf. Sci. Technol., vol. 59, no. 4, pp. 526–534, 2008.

[5] A. Martín-Martín, E. Orduña-Malea, J. M. Ayllón, and E. D. López-Cózar, “Back to the past: on the shoulders of an academic search engine giant,” Scientometrics, vol. 107, no. 3, pp. 1477–1487, 2016.

[6] T. Tonia, H. Van Oyen, A. Berger, C. Schindler, and N. Künzli, “If I tweet will you cite? The effect of social media exposure of articles on downloads and citations,” Int. J. Public Health, vol. 61, no. 4, pp. 513–520, May 2016.

[7] S. J. Cunningham and D. Bocock, “Obsolescence of computing literature,” Scientometrics, vol. 34, no. 2, pp. 255–262, 1995.

[8] B. F. Jones and B. A. Weinberg, “Age dynamics in

scientific creativity," Proc. Natl. Acad. Sci. U. S. A., vol. 108, no. 47, pp. 18910–18914, Nov. 2011.

[9] M. Šember, L. Škorić, and J. Petrak, "Current impact of ceased journals: are they still alive?," Malaysian Journal of Library & Information Science, vol. 22, no. 1, pp. 15–27, 2017.

[10] C. Oppenheim and S. P. Renn, "Highly cited old papers and the reasons why they continue to be cited," Journal of the American Society for Information Science, vol. 29, no. 5, pp. 225–231, 1978.

[11] L. Zhang and W. Glänzel, "A citation-based cross-disciplinary study on literature aging: part I—the synchronous approach," Scientometrics, vol. 111, no. 3, pp. 1573–1589, 2017.

[12] L. Egghe, I. K. Ravichandra Rao, and R. Rousseau, "On the influence of production on utilization functions: Obsolescence or increased use?," Scientometrics, vol. 34, no. 2, pp. 285–315, 1995.

[13] W. Glänzel, B. Thijs, and P.-S. Chi, "The challenges to expand bibliometric studies from periodical literature to monographic literature with a new data source: the book citation index," Scientometrics, vol. 109, no. 3, pp. 2165–2179, 2016.

[14] M. Thelwall and T. Nevill, "Could scientists use Altmetric.com scores to predict longer term citation counts?," J. Informetr., vol. 12, no. 1, pp. 237–248, 2018.

[15] D. I. Stern, "High-ranked social science journal articles can be identified from early citation information," PLoS One, vol. 9, no. 11, p. e112520, Nov. 2014.

[16] Brookes, B. C. "Numerical Methods of Bibliographic Analysis", Library Trer~Is 22:8; 1973.

[17] D. P. Wallace, The Relationship Between Journal Productivity and Obsolescence in a Subject Literature. 1985.

[18] S. Sikdar, M. Marsili, N. Ganguly, and A. Mukherjee, "Influence of Reviewer Interaction Network on Long-Term Citations: A Case Study of the Scientific Peer-Review System of the Journal of High Energy Physics," in 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2017.

[19] J. Sun, C. Min, and J. Li, "A vector for measuring obsolescence of scientific articles," Scientometrics, vol. 107, no. 2, pp. 745–757, 2016.

[20] C. Nansen and W. G. Meikle, "Journal impact factors and the influence of age and number of citations," Mol. Plant Pathol., vol. 15, no. 3, pp. 223–225, Apr. 2014.

[21] V. Larivière, E Archambault and Y Gingras "Long-Term Variations in the Aging of Scientific Literature: From Exponential Growth to Steady-State Science (1900–2004)"

[22] Luwel, M. & Moed, H.F., "Publication delays in the science field and their relationship to the ageing of scientific literature. Scientometrics" (1998) 41: 29

[23] Phil Davis, "Why are Authors Citing Older Papers?" APR 29, 2015

[24] Hajra and Sen, "Aging in citation networks", Physica A: Statistical Mechanics and its Applications Volume 346, Issues 1–2, 1 February 2005, Pages 44-48

[25] L. Egghe and R. Rousseau, "Aging, obsolescence, impact, growth, and utilization: Definitions and relations", Journal of the Association for Information Science and Technology, Volume 51, Issue 11, Pages: 969-1066

[26] M.V. Simkin and V.P. Roychowdhury, "A mathematical theory of citing", Journal of the Association for Information Science and Technology, Volume 58, Issue 11, Pages: 1557-1706

[27] EGGHE, Leo(2001) "A non-informetric analysis of the relationship between citation age and journal productivity" Journal of the American Society for Information Science and Technology, 52(5). p. 371-377

[28] D.P. Wallace (1986). "The relationship between journal productivity and obsolescence". Journal of the American Society for Information Science 37(3), 136-145

[29] Y. Yao, H. Tong, F. Xu and J. Lu, "Predicting long-term impact of CQA posts: a comprehensive viewpoint", August 2014 KDD '14: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.

[30] C. Stegehuis, N. Litvak and L. Waltman, "Predicting the long-term citation impact of recent publications", Journal of Informetrics, Volume 9, Issue 3, July 2015, Pages 642-657

[31] D. Wang, C. Song and A. Barabasi, "Quantifying Long-Term Scientific Impact", Science 04 Oct 2013, Vol. 342, Issue 6154, pp. 127-132.

[32] S. Redner, How popular is your paper? An empirical study of the citation distribution, Eur. Phys. J. B, 4 (1998), pp. 131–134.

[33] S. Datta, S. Sarkar, and A. S. M. Sajeev. How long will this live? discovering the lifespans of software engineering ideas. IEEE Transactions on Big Data, 2(2):124–137, jun 2016.

[34] Mayank Singh, Ajay Jaiswal, Priya Shree, Arindam Pal, Animesh Mukherjee, and Pawan Goyal. 2017. Understanding the Impact of Early Citers on Long-Term Scientific Impact. In Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on. IEEE, 1–10.

[35] E. Matricciani, "Shannon's entropy as a measure of

the ‘life’ of the literature of a discipline”, Scientometrics, May 1994, Volume 30, Issue 1, pp 129–145