Asian Journal of Information Technology 15 (12): 1964-1969, 2016
ISSN: 1682-3915

Discovering Citation Manipulations via Delayed Citation Recognition

[1]R. Siva, [2]G.S. Mahalakshmi and [3]S. Sendhilkumar
[1]Department of Computer Science and Engineering,
KCG College of Technology, Karappakam, 600097 Chennai, India
[2]Department of Computer Science and Engineering, Anna University,
600025 Chennai, India
[3]Department of Information Science and Technology, Anna University,
600025 Chennai, India

Abstract: In this study, the citations obtained for a published paper are analyzed to find the citation behavior. The published study is compared with their respective citations to find the research utilization factor. This research utilization factor is based on the proposedcitation function. Through this research, an attempt is made to find the similarity in the pattern of citation utilization across delayed and non-delayed cited study.

Key words: Citation behavior, delayed cited study, utilization factor, function, pattern

INTRODUCTION

In this study, the citation behavior of the published paper is analyzed and an attempt is made to find the pattern for the citation behavior of a particular paper. The published paper might get the citation on the same year it has published; such papers are considered to be non-delayed papers. The published papers that get citation only after few years of the published year are considered to be delayed recognized paper. The published paper along with the all the cited papers are identified. Then, these papers are analyzed to find the citation purpose and citation behavior. The citation behavior is analyzed to be categorized into twelve parts which is based on the citation function of Simon Teufal Classification. These twelve categories has a particular arrangement which has been experimentally studied and most of the non-delayed papers seem to satisfy this particular arrangement of citation. But the delayed recognized paper seems to be deviating from this particular citation behavior. Thus, this paper provides an experimental proof that the citation behavior of delayed and non delayed papers are not similar.

Previous studies have studied about the citation behavior based on multiple approaches. In the work by Amjad Abu-Jbara on Reference Scope Identification in Citing Sentences they have considered that every cited paper will have a sentence about the corresponding base paper which they refer as the citing sentence. This citing sentences contains the purpose of utilizing the base paper and other related details. So, they have identified this citing sentence and have analyzed to find the reason of citation Awais Athar and this is further user for identification of the scope of the reference.

In the study by Athar and Teufel (2012) on identification of citation function, an attempt was made to find all the possibilities of citing or referring a paper. There were twelve citation functions which are identified as the citation reasons. Then, the paper and cited paper are analyzed to find the reason or purpose behind citation. On analysis they have found out that the purpose can be categorized into three categories, weak, neutral and positive. Based on these categories they have analyzed the cited paper to fall under any of the one categories. In the work by Lachance and Lariviere (2014), the concept of delayed cited paper is introduced. The delayed papers are the published papers which get the recognition only after few years from publication. Such papers are identified and the citation behaviors of such delayed papers are studied. Finally, the citation lifecycle of the delayed paper is found out and studied. Research progress analysis (Balaji *et al*., 2016) has been proposed as interesting research problem in bibliometrics research. Citations serve a prominent role in analysing an individual's scientific research progress. However, manipulation in citations remains a challenge in real root cause analysis.

MATERIALS AND METHODS

The published paper which gets its first citation within the year of publication is identified as the non-delayed paper. The published paper which gets its

Corresponding Author: R. Siva, Department of Computer Science and Engineering, KCG College of Technology, Karappakam, 600097 Chennai, India

first citation after one or more year from the year of publication is identified as delayed paper. The data set consists of research papers, published across years. The published papers across their citations are analysed for identifying if it is under delayed or non-delayed category. This identification of delayed and non-delayed is achieved by comparing the published years. If the base paper's published year and any one of the cited paper's published year are same then that base paper can be separated as non-delayed category. If there doesn't exist a cited paper, with the same publication year as the base paper's publication year, then such papers are identified and separated as the delayed category. After separating the delayed and non-delayed categories, further analysis is madeover the citation behavior of paper in respective categories.

Citation behavior of non-delayed paper: First, the non-delayed papers, i.e., the papers that got immediate citation are taken for analysis. Here, in this analysis we are trying to find the order in which the work in the base paper might be utilized by the cited papers. This utility order can be found by finding the utilization criteria of each cited paper with the base paper. Thus, the non-delayed paper is compared with each of it's cited paper to find the utilization criteria. For example, if a non-delayed base paper has fourteen citations, then these 14 cited papers will be individually compared with the base paper to find the utilization criteria. Then based on the utilization criteria specific order for the citation behavior of non-delayed paper can be found.

Arrangement of cited paper in the order of cited year: In order to find the utilization order, the cited papers needs to be arranged in the order of the year. The study which made the first citation will have different utilization criteria, compared to the last paper which cited the base paper recently. Hence, it is highly important to arrange the cited paper of the base paper in the order of the year they cited the base paper. The cited paper, thus arranged according to their year of citation is compared with the base paper to find the utilization criteria.

Cue phrase identification: Cue phrases Rashid and Teufel are meta-discourse is used to identify the work in the paper. There is a set of phrases or verbs that are often used to find the entire goal of the paper. There are certain verbs like suggest, describe, based on etc. which describes the base utilization. Verbs like apply, based on, be originated in etc., can be used to find use utilization. There are also specialized verb clusters to which co-occur with base, e.g., the cluster of continuation of ideas (e.g., "adopt, agree with, base, be based on, be derived from, be originated in be inspired by, borrow, build on, etc). Weak sentences is often concerned with failing (of other researchers' approaches) and often contain verbs such as abound, aggravate, arise, be cursed, be incapable of be forced to be limited to etc. Similarly each utilization are assigned with the cue phrase and verb in order to find the similarity (Zhu *et al*., 2011) between the base paper and the cited paper. A total of 20 cue phrases are used to find the utilization factor by the cited paper.

Comparison based on similarity measure: Apart from this method of using cue phrases, the similarity measure between the base paper and the cited paper is also performed (Martin *et al*., 2011, 2013). The summarized version of both the base paper and the cited paper is semantically analyzed to find semantic structure of documents, by examining word statistical co-occurrence patterns within a corpus. This corpus collection can beused to automatically infer structure of the documents, their topics etc. The corpus also includes the cue phrases mentioned above. Then vector space model is constructed which gives the frequency of the corpus between the base paper and the cited paper. This vector representation of the document can be used to infer the utilization criteria of the cited paper and thereby will help in analyzing the citation graph of every published research manuscript for correctness (Mahalakshmi and Sendhilkumar, 2013; Lu *et al*., 2012; Medina and Noyons, 2008).

Order of utilization in citation: Based on the inference obtained on comparing the non-delayed base paper and its corresponding cited papers which are arranged in the order of year of citation it has been found that there is a particular order in which the base paper is being used by the cited papers. This order of utilization interestingly follows the theme of systematic research which a published paper undergoes in the research journey amidst the peer works in the respective research arena. In the initial years of publication of base paper, the cited paper used the base paper for the purpose of utilizing the work for their work. As few years passes, the work in base paper is modified or contradicted by the cited paper. Again, few years later, the work in base paper has been just cited or used for efficiency comparison by the cited paper. Therefore, we propose the order of citation behavior of the non-delayed paper as: base, use, similarity, motivate, support, neutral, modify, compare, contrast, weak, methods, result. The point to note here is that the non-delayed papers are the research papers which received due timely citation recognition in the respective research disciplines and hence are said to follow the cycle of citation utilization (Table 1).

Table 1: Citation utilisation cycle of non-delayed research recognition

Order	Utilization criteria	Description
0	Base	Researcher uses cited work as basis or starting point
1	Use	Researcher uses tools/algorithms
2	Similarity	Researcher's work and cited work are similar
3	Motivate	Researcher is positive about the approach used or problem addressed
4	Support/enhance	Little contribution to base work
5	Neutral	Neutral description of cited work
6	Modify	Researcher adopts new algorithm from existing base work
7	Compare	Comparing researcher's work and cited work and other existing works
8	Contradict	Contradict the approach adopted by the base work
9	Weak	No or less similarity (just cited)
10	Methods	Comparing methods and approaches
11	Results	Comparing the end result

Annotation utilisation criteria in citation behavior: The 1st stage of citation behavior is base. In this stage, the paper will be newly published and hence the initial paper citing this study will use this paper as the base for their work. The cited work will be used by the author as the basis or starting point for author's work. Therefore, the cited work will be used for the study and as the first step for further work extension.

The 2nd stage of citation behavior is use. The cited paper would have proposed some new tools or algorithms or maybe a new approach for a problem statement. These algorithms or tools will be used by the author for his work which could be related to the same problem statement as the cited work. Hence, the author would have cited the base paper for using the same concept in his work.

The 3rd stage of citation behavior is similarity. The cited paper would have the work which is similar to the work that the author is working on. To get the idea or the approach the cited paper has adopted, the author would cite the base paper. Hence, the final contribution of the author who cited is going to be similar to the cited work.

The 4th stage of citation behavior is Motivate. The researcher citing the base study would appreciate or motivate the work in base study. The author would motivate that the method or approach adopted to solve the particular problem is right and claims that the problem can be solved only by the approach proposed in cited work.

The 5th stage of citation behavior is support/enhance. The researcher citing the base paper would enhance the cited work by contributing a little more work to the existing work in the base paper. The contribution done by the cited author would be in such a way that it enhances the existing work for better performance.

The 6th stage of citation behavior is neutral. The work done in the base paper will not be appraised or dispraised. The author would have just given some credit to the base paper for doing some relevant work on the chosen problem statement. Thus, the cited author would be neutral about the approach handled by the base researcher.

The 7th stage of citation behavior is modify. The cited researcher would modify the entire approach or algorithm either to handle every cases of the problem statement or to make the algorithm efficient. This modification can be improvisation of the existing approach. There are few cases of changing the entire approach which also comes under this modification.

The 8th stage of citation behavior is compare. The cited author would have gone through many solutions of the same problem statement. The cited author would give a comparison account of each and every approach adopted to solve the particular problem. In some cases, along with the comparison account, the cited author would suggest a better approach which would perform better compared to the approach in base work. Such papers are classified as compare.

The 9th stage of citation behavior is contrast. The researcher who cited base work would contradict the approach used in the base work. The cited author would contradict that the entire approach must be handled differently, not as suggested in the base work. The cited author would find flaws in the approach and the remedies for that flaw might or might not be handled by the cited reseacher.

The 10th stage of citation behavior is weak. After being contradicted, the author would just cite the base paper in order to get the idea of what are the problems faced by the approach adopted by the base work. Thus, the cited author would just mention the paper in the reference. The work in the base would not be utilized, leading to the weakness of the paper.

The 11th stage of citation behavior is method. In this case, the researcher would have cited the base paper for the sake of comparison of the methods that are existing. The method comparison will be used for finding and comparing efficiency of the approach adopted by cited author.

The 12th stage of citation behavior is result. The results of the base paper will be compared with the cited researcher's end result. If the result of the cited author's work is more accurate then the cited author would use the end result comparison to show the superiority of the work done by the cited researcher.

Citation behavior of delayed paper: The study which gets citation after few years of publication is considered as the delayed paper. The delayed paper and its cited paper are analyzed in the same way as the non-delayed paper. The

utilization criteria of the delayed cited paper is found out. It has been found that the order in which non-delayed papers are cited is not the same for the delayed paper. There seemed to be some manipulation or repeated utilization for the same purpose. This has been experimentally proved for the dataset consisting research papers from the journal scientometrics.

RESULTS AND DISCUSSION

The dataset consist of published research studies from the journal scientometrics, springer, between 1997-2007. There were around 250 papers in each year. These stueies are split into delayed and non-delayed paper, based on the comparison of first cited paper year. Then, in each year the paper with citation >14 are consider for evaluation in order to find citation behavior.

Let us say, the delayed paper be represented as follows, delayed paper, DP = $\{X_1, X_2, X_3, ..., X_n\}$ where $\{X_1, X_2, X_3, ..., X_n\}$ are the set of research papers that acquired delayed citation. Let us say, the non-delayed paper be represented as follows, NDP = $\{Y_1, Y_2, Y_3, ..., Y_n\}$ where $\{Y_1, Y_2, Y_3, ..., Y_n\}$ are the set of research studies which acquired immediate citation.

Each paper in delayed and non delayed set has group of cited papers which are represented as follows, $X_i = \{x_{i1}, x_{i2}, x_{i3}, ..., x_{im}\}$ where $\{x_{i1}, x_{i2}, x_{i3}, ..., x_{im}\}$ is the citation list of X_i and $Y_i = \{y_{i1}, y_{i2}, y_{i3}, ..., y_{im}\}$ where $\{y_{i1}, y_{i2}, y_{i3}, ..., y_{im}\}$ is the citation list of Y_i, respectively.

The cited papers of each paper are grouped in a pair of two in the order of the cited year. This grouping is done because there are papers whose time difference between two citations may not be same. In order to generalize this the grouping is done in the order of citation. Citation grouping of delayed paper $DCG_{ij} = \{x_{i(j-1)}, x_{ij}\}$ where $x_{i(j-1)}$, x_{ij} are $(j-1)^{th}$ and jth cited papers of base delayed paper X_i. Citation grouping of non-delayed paper $NDCG_{ij} = \{y_{i(j-1)}, y_{ij}\}$ where $y_{i(j-1)}$, y_{ij} are $(j-1)^{th}$ and jth cited papers of base non-delayed paper Y_i.

In each citation grouping, the overall maximum value of each citation group is considered to be the citation purpose or utilization value of that particular group. Maximum similarity value in citation grouping of delayed paper is:

$$DCG_i = \sum_{j=1}^{m} \max[x_{i(j-1)}, x_{ij}]$$

where, max $[x_{i(j-1)}, x_{ij}]$ gives the maximum of utilization value of $x_{i(j-1)}$ and x_{ij} cited paper of base delayed paper X_i. Maximum similarity value in citation grouping of non-delayed paper is:

$$NDCG_i = \sum_{j=1}^{m} \max [y_{i(j-1)}, y_{ij}]$$

where, max $[y_{i(j-1)}, y_{ij}]$ gives the maximum of utilization value of $y_{i(j-1)}$ and cited paper of base non-delayed paper Y_i.

The maximum utilization value for each citation grouping per paper is averaged to find the overall utilization value Cud and thereby the citation behavior:

$$C_{ud} = \left(\frac{\sum_{i=1}^{n} DCG_i}{n} \right)$$

The maximum utilization value for pth citation group pair for all delayed papers X_i are averaged to find the overall utilization value C_{ud_p}:

$$C_{ud} = \left(\frac{\sum_{i=1}^{n} [x_{i(p-1)}, x_{ip}]}{n} \right)$$

where, $p \leq m$, p is the position of citation group pairs, k and m is the total number of citation group pairs for X_i. These average values are plotted against citation utilization criteria value in a line graph to find the citation behavior. The graph obtained is a non-straight graph as shown in Fig. 1 and Table 2.

The maximum utilization value for each citation grouping per paper is averaged to find the overall utilization (aka early utilization) value C_{ud} and thereby the citation behavior.

The maximum utilization value for pth citation group pair for all non-delayed papers Y_i are averaged to find the overall utilization value C_{ud_p}:

$$C_{ud_p} = \left(\frac{\sum_{i=1}^{n} \max[x_{i(p-1)}, y_{ip}]}{n} \right)$$

where, $p \leq m$, p is the position of citation group pairs, k and m is the total number of citation group pairs for Y_i. These average values are plotted against citation utilization criteria value in a line graph to find the citation behavior. The graph obtained is a non-straight graph as shown in Fig. 1 and Table 3.

Figure 2 describes the delayed paper citation behavior. A graph is plotted with cited paper grouping and the utilization criteria. It has been found that

Table 2: Delayed study citation utitlization

	C_{ud}						
AVG	1	2	3	4	5	6	7
Utilization value	0.5	2.9	6	3.4	2.3	7	9.1

Table 3: Non-delayed study citation utilization value

AVG	A1	A2	A3	A4	A5	A6	A7
Utilizationvalue	0.4	1.9	4.4	6.8	7.4	8.3	9.5

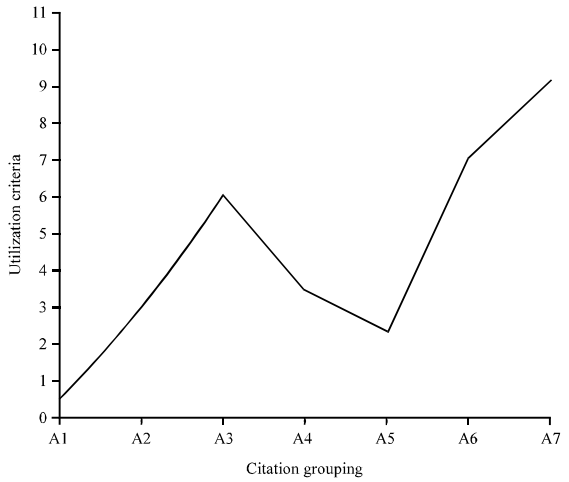

Fig. 1: Delayed paper citation behavior based on utilization value

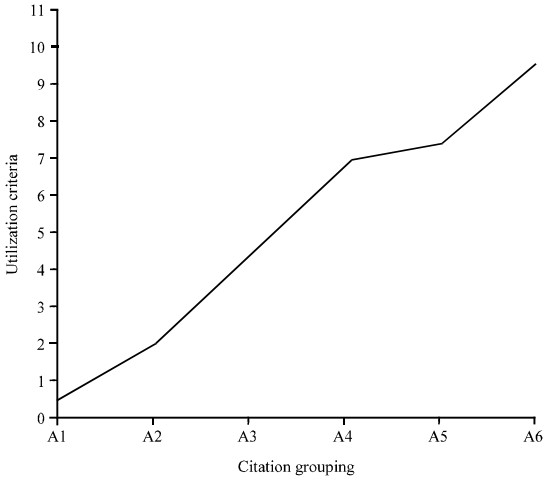

Fig. 2: Non-delayed paper citation behavior based on utilization value

the non-delayed papers follow the order of utilization criteria that has been found experimentally. The line graph shows the linear behavior which proves that the citation behavior occurs in the order of utilization criteria. Figure 2 describes the non-delayed paper citation behavior. The cited paper grouping and the utilization criteria are plotted in the line graph. The line graph thus obtained shows a non-linear behavior which denotes that the order of utilization criteria is not the same for delayed paper. There is some repetitive utilization in the non-delayed paper.

The behavior of non-delayed papers approximately obey the proposed order of research utlisation criteria whereas the behavior of delayed category reveals that the papers do not comprise to the order or research purpose. Figure 1 shows a reversal of A5 which the average of firth citation group pairs of all delayed papers which is close to that of first citation group average. In other words as the research progresses in the direction of modify, parallely there was some gap spotted in the base paper which was left untouched in terms of research continuity and therefore, such parts of the base paper appears visible to the citees very lately and therefore, the journey of research utilization has repeated from the start. This indirectly indicates either there is a lack of clarity in the writing of base paper or in other words, the base paper was overlooked earlier and therefore manipulated until there comes a real necessity for citing the base paper arised.

CONCLUSION

The dataset used is 3000+ research papers from Journal Scientometrics, Springer Publishing between 1997 and 2007. The citation histories of delayed and non-delayed paper of this dataset arecompared. It can be observed from the result that the pattern followed by the non-delayed paper on citation behavior is not the same for the delayed paper. Almost, every non-delayed paper obey the order of utilization criteria found by experimental results. Whereas the delayed paper doesn't follow the order of utilization criteria and the utilization seems to be the same as the citation increases. This repeated utilization maybe because of the self citation or manipulation in the citation by the author who published. Hence, this work can be extended to find this abnormal citation behavior of the delayed paper. The reason for repeated citation utilization of the delayed paper can be found by closely analyzing the delayed paper's cited paper.

REFERENCES

Athar, A. and S. Teufel, 2012. Context-enhanced citation sentiment detection. Proceedings of the Conference on North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 3-8, 2012, Association for Computational Linguistics, Stroudsburg, Pennsylvania, USA., ISBN: 978-1-937284-20-6, pp: 597-601.

Balaji, A,. S. Sendhilkumar and G.S. Mahalakshmi, 2016. Progressive path analysis using optimized discrete and continuous average semantic filters. Aust. J. Basic Appl. Sci., 10: 224-233.

Lachance, C. and V. Lariviere, 2014. On the citation lifecycle of papers with delayed recognition. J. Inf., 8: 863-872.

Lu, L.Y.Y., Y.L. Lan and J.S. Liu, 2012. A novel approach for exploring technological development trajectories. Proceedings of the International Conference on Management of Innovation and Technology, June 11-13, 2012, IEEE, Sanur Bali, ISBN: 978-1-4673-0108-4, pp: 504-509.

Mahalakshmi, G.S. and S. Sendhilkumar, 2013. Optimizing research progress trajectories with semantic power graphs. In: Pattern Recognition and Machine Intelligence. Maji, P., A. Ghosh, M.N. Murty, K. Ghosh and S.K. Pal (Eds.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-642-45061-7, pp: 708.

Martin, G.H., S. Schockaert, C. Cornelis and H. Naessens, 2011. Finding similar research papers using language models. Proceedings of the 2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation, October 23-24, 2011, University College Ghent, Bonn, Germany, pp: 106-113.

Martin, G.H., S. Schockaert, C. Cornelis and H. Naessens, 2013. Using semi-structured data for assessing research paper similarity. Inf. Sci., 221: 245-261.

Medina, C.C. and E.C. Noyons, 2008. Combining mapping and citation network analysis for a better understanding of the scientific development: The case of the absorptive capacity field. J. Inf., 2: 272-279.

Zhu, S., J. Wu, H. Xiong and G. Xia, 2011. Scaling up top-K cosine similarity search. Data Knowl. Eng., 70: 60-83.