Citation sentence reuse behavior of scientists: A case study on massive bibliographic text dataset of computer science

Mayank Singh
Dept. of Computer Science and Engg.
IIT Kharagpur, India
mayank.singh@cse.iitkgp.ernet.in

Abhishek Niranjan
Dept. of Computer Science and Engg.
IIT Kharagpur, India
aniranjan@cse.iitkgp.ernet.in

Divyansh Gupta
Dept. of Computer Science and Engg.
IIT Kharagpur, India
divyansh.gupta@cse.iitkgp.ernet.in

Nikhil Angad Bakshi
Dept. of Mechanical Engg.
IIT Kharagpur, India
nabakshi@iitkgp.ac.in

Animesh Mukherjee
Dept. of Computer Science and Engg.
IIT Kharagpur, India
animeshm@cse.iitkgp.ernet.in

Pawan Goyal
Dept. of Computer Science and Engg.
IIT Kharagpur, India
pawang@cse.iitkgp.ernet.in

ABSTRACT

Our current knowledge of scholarly plagiarism is largely based on the similarity between full text research articles. In this paper, we propose an innovative and novel conceptualization of scholarly plagiarism in the form of reuse of explicit citation sentences in scientific research articles. Note that while full-text plagiarism is an indicator of a gross-level behavior, copying of citation sentences is a more nuanced micro-scale phenomenon observed even for well-known researchers. The current work poses several interesting questions and attempts to answer them by empirically investigating a large bibliographic text dataset from computer science containing millions of lines of citation sentences. In particular, we report evidences of massive copying behavior. We also present several striking real examples throughout the paper to showcase widespread adoption of this undesirable practice. In contrast to the popular perception, we find that copying tendency increases as an author matures. The copying behavior is reported to exist in all fields of computer science; however, the theoretical fields indicate more copying than the applied fields.

CCS CONCEPTS

•Information systems →Near-duplicate and plagiarism detection;

KEYWORDS

Citation context, plagiarism, text reuse

ACM Reference format:
Mayank Singh, Abhishek Niranjan, Divyansh Gupta, Nikhil Angad Bakshi, Animesh Mukherjee, and Pawan Goyal. 2017. Citation sentence reuse behavior of scientists: A case study on
massive bibliographic text dataset of computer science. In *Proceedings of The ACM/IEEE-CS Joint Conference on Digital Libraries, Toronto, Ontario, Canada, June 2017 (JCDL '17),* 5 pages.
DOI: 10.475/1234

JCDL '17, Toronto, Ontario, Canada

DOI: 10.475/1234

1 INTRODUCTION

Scholarly text "reuse" detection is a well known problem. It has received even more attention in the past decade due to overwhelming increase in the literature volume and ever increasing cases of plagiarism [9]. Traditionally, the focus has always been in the analysis of full text of the research articles. A highly celebrated work on copy detection by Brin et al. [5] proposed a working system, called COPS, that detects copies (complete or partial) of research articles. They also proposed algorithms and metrics required for evaluating detection mechanisms. Bao et al. [3] applied Semantic Sequence Kin for document copy detection. Zu et al. [17] tried to detect plagiarism if references are not given. They develop a taxonomy of plagiarism delicts along with features for the quantification of style aspects. Recently, Citron et al. [7] described three classes of text reuse. They studied text reuse via a systematic pairwise comparison of the text content of all articles submitted to arXiv.org between 1991 – 2012. They report that in some countries 15% of submissions are detected as containing duplicated material. Lesk [10] presented scope of plagiarism within arXiv. They concluded that arXiv is now identifying the papers that have substantial overlap and is waiting to see if that affects the submission.

In recent past, numerous cases of widespread plagiarism have been detected leading to severe consequences. For example, Professor Matthew Whitaker from Arizona State University was made to resign after a series of plagiarism controversies[1] . Professor Shahid Azam from University of Regina has been accused for plagiarizing his own student master's thesis [2]. Leading to an even more disastrous consequence, a city court in India briefly sent former vice-chancellor to jail on allegations of plagiarism [2].

Plagiarism detection is considered computationally demanding as majority of proposed techniques rely on similarity between full text research articles. In this paper, we propose a more nuanced micro-scale copying phenomenon by examining crowdsourced data generated in the form of explicit citation sentences. In contrast to full-text plagiarism that exhibits a gross-level behavior, we present a first attempt to understand the copying of citation sentences that corresponds to a more nuanced micro-scale phenomenon. We believe that researchers should be capable to describe previous literature without plagiarizing. To improve the quality and innovation, scientific community should strongly discourage such activities.

[1]Wikipedia article: https://en.wikipedia.org/wiki/Matthew_C._Whitaker#Controversy

2 DATASET

In this paper, we use two computer science datasets crawled from Microsoft Academic Search (MAS)[2]. The first dataset [6] consists of bibliographic information (the title, the abstract, the keywords, its author(s), the year of publication, the publication venue, and the references) of more than 2.4 million papers published between 1859 – 2012. The second dataset [14] consists of more than 26 million citation sentences present all across the computer science articles published in the same time window. The scripts and processed data is available online[3] for download.

3 CITATION SENTENCES AND SIMILARITY

Throughout this paper we use the terms 'citation sentence' and 'citation context' interchangeably. If paper P refers to paper C, then P is termed the citing paper while C is termed the cited paper. Given paper P, we consider those sentences as *citation sentences* (C_S) that *explicitly* cite the previous paper C. Note that, P can refer to C at many places in the text leading to multiple C_S for the same cited-citer pair. We process raw C_S by replacing all citation placeholders (reference indexes, author names plus year etc.) with a single word "CITATION". We have been successful in replacing 16 different citation placeholder formats identified by Singh et al. [13]. A representative citation context from our dataset where [1] cites [8], before and after pre-processing is as follows:

Before: *Recommender systems are a personalized information filtering technology [4], designed to assist users in locating items of interest by providing useful recommendations.*

After: *Recommender systems are a personalized information filtering technology CITATION, designed to assist users in locating items of interest by providing useful recommendations.*

Similarity computation: This study massively relies on similarity scores between two citation contexts. Therefore, we utilize vector space model to compute similarity scores using tf-idf weighting scheme. We construct vocabulary from rich scientific text present in the second dataset. Due to computation complexity associated with similarity computations, out of ~1.5 million tokens, we only consider top 100,000 frequently occurring tokens. For each pair of C_S vectors, similarity scores are generated using standard cosine similarity metric (*CosSim*). We employ python's machine learning library scikit-learn[4] for all the computations.

4 MOTIVATIONAL STATISTICS

We begin this work by examining the most intriguing and trivial question – *Whether* C_S are really copied and to what extent? To motivate the reader, we present a representative example of extensive copying from our dataset. Later in this section, we demonstrate that copying behavior is not a rare phenomenon; in contrast, it seems to have become a widespread convention.

A representative example: We found a large number of articles in our dataset whose incoming C_S were partially or completely copied. In particular, we found cases where a paper receives an exact copy of C_S from different citing papers. As a representative example, we found five copies of the citation context – *"We do not aim at formalizing some specific kind of state diagram, which has already successfully been done, see CITATION for example"*. Here "CITATION" placeholder consists of seven cited papers.

Overall copying behavior: Further, we attempted to understand the overall C_S copying behavior. To start with, we randomly selected 24,800 papers. For each paper p, we compute pairwise cosine similarity between the incoming C_S. Figure 1 presents the distribution of similarity scores. As expected, most C_S pairs have very low similarity scores ($CosSim \leq 0.2$). However, we also observe significant number of pairs having high similarity ($CosSim \geq 0.8$). Most surprisingly, we found a sharp peak at $CosSim = 1$; highlighting a very interesting observation that many C_S are directly copied without any change. We also found papers where C_S are copied from multiple articles.

Overall, we found ~ 26 thousand articles that consists of at least one citation context ditto copied ($CosSim = 1$) from another paper. We have used a strict metric for this; in specific, we concatenate the multiple instances for C_S between the same pair of papers so that every pair has a unique citation context. Among this set, we find 148 articles each having at least five C_S, with $\geq$ 60% of C_S being exact copy from previously published papers.

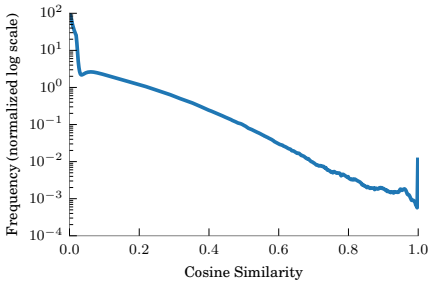

Figure 1: Cosine similarity score distribution for C_S pairs.

These initial statistics offer compelling evidence to study the phenomenon of copying C_S in-depth. In the next section, we present extensive empirical analysis to understand the effect and properties of this interesting albeit undesirable phenomenon.

5 LARGE SCALE EMPIRICAL STUDY

In this section, we pose several interesting questions to better understand the characteristics of this micro-scale copying behavior. To answer these questions, we conduct in-depth empirical analysis on the computer science dataset described in section 2. Note that due to associated computation complexities in processing large text data, we perform individual experiments on smaller samples[5] of full dataset. For better interpretation, the posed questions are grouped into three categories: 1) Paperwise, 2) Authorwise, and 3) Fieldwise. We consider copying if pairwise cosine similarity between incoming C_S is higher than 0.8.

5.1 Paperwise analysis

5.1.1 Does publication age impact copying behavior? In this section, we attempt to investigate the temporal nature of the copying behavior. To start with, we select top 500 cited papers (P). For each selected paper $p \in P$, we study its incoming C_S (that describe p) within Δt years after publication. For the current study, we use

[2] http://academic.research.microsoft.com
[3] https://tinyurl.com/kzv4lhg
[4] http://scikit-learn.org/stable/
[5] We describe sample statistics in the beginning of each experiment.

$\Delta t = 1, 2, 3, 4, 5$. Evidence suggests that copying behavior changes with the passage of time. More specifically, it first decreases and then starts increasing over the time-period with the exception for $\Delta t = 1$. In each $y \in \Delta t$, we compute the fraction of C_S copied (denoted by f_q) for each of the paper q, which cites paper $p \in P$. We compute the fraction of copied C_S at Δt year after publication by $F(\Delta t) = \frac{\sum_{q=1}^{n} f_q}{n}$. For each Δt, we compare similarity between C_S generated in the year Δt with all the C_S generated between $\Delta(t = 0)$ to $\Delta(t - 1)$. We find the value of average fraction of copied C_S = 8.33% in the first year after publication. In the next four successive years, the fraction varies as follows, 8.97%, 8.02%, 8.52%, 9.12%, indicating that the copying behavior decreases first and then gradually increases.

5.1.2 Are there differences in the cited versus the non-cited copying? Next, we divide copied C_S into two subsets, namely, i) cited (*CC*), and ii) not cited (*NC*). *CC* consists of copied C_S where the source paper (from which context is being copied) is cited, whereas *NC* consists of copied C_S where the source paper is not cited. Figure 2a reports proportion of copied C_S into these subsets at five time periods after publication. To our surprise, we find that fraction of two subsets remains same over the years. Fraction of *NC* is significantly lower than fraction of *CC*.

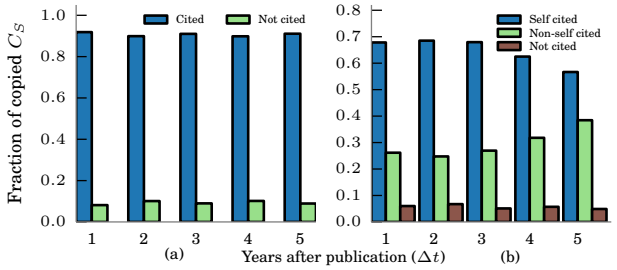

Figure 2: a) Fraction of copied C_S in two subsets, namely, cited and not cited. The fraction of two subsets remains nearly the same over the years. b) Fraction of C_S in three subsets, namely, self cited, non-self cited and not cited. The fraction of *self cited* subset decline over the years leading to increase in the fraction of *non-self cited* subset.

The copy and cited behavior is a combination of two distinct forms of copying, namely, *self copying* and *non-self copying*. In self copying, an author copies her own older C_S, whereas in non-self copying, an author copies C_S written by other authors. Therefore, we split *CC* subset into further two subsets, namely, i) self cited (*SC*), and ii) non-self cited (*NSC*). Figure 2b reports proportion of copied C_S into three subsets (*SC*, *NSC* and *NC*) at five time periods after publication. The single most striking observation from Figure 2b is that fraction of *SC* is much higher than *NSC* indicating that majority of the C_S are copied by an author in their own future publications. However, as authors' interest shifts from one topic to other, the fraction of *SC* declines resulting into increase in the fraction of *NSC*.

5.2 Authorwise analysis

5.2.1 Does an author's increasing experience influence her copying behavior? In this section, we present empirical results to prove that researchers in their early stage of academic career behave differently than after gaining experience. We begin this analysis by selecting 7175 random authors that have at least 20 years of citation history. For each author, we compute fraction of copied C_S from previously published papers. Figure 3a presents average fraction of copied C_S over 20 years of the author life span. In contrast to the popular perception, we observe that the copying tendency increases as an author matures.

5.2.2 Does an author's popularity influence his copying behavior? We observe that an author's popularity plays a critical role in influencing his copying patterns. For this study, we select authors with varying popularity but with similar academic age[6]. We select all authors that started their career from the year 1995. We compute the authors' h-index (in 2012) to measure their individual popularity. To better visualize the influencing behavior, we divide the authors into three h-index buckets:

- **Bucket 1:** h-index $<$ 5
- **Bucket 2:** 5 $\leq$ h-index $<$ 15
- **Bucket 3:** h-index $\geq$ 15

Here **Bucket 1** represents the least popular authors whereas **Bucket 3** consists of the most popular authors. We compute the fraction of copied C_S for each author in each bucket and present aggregated statistics. We observe that the most popular authors (**Bucket 3**) show maximum copying tendency. Whereas least popular authors (**Bucket 1**) show least copying tendency. On average, 2.07% C_S of **Bucket 1** are copied from previous papers. This fraction increases to 4.17% and 6.16% for **Bucket 2** and **Bucket 3** respectively. To investigate in more detail, we divide copied C_S into three subsets as described in section 5.1.1. Figure 3b presents proportion for three different copying behaviors in three h-index buckets. Fraction of *SC* increases as h-index increases. Popular authors prefer to copy their own C_S while less popular authors try to copy C_S written by other authors. Figure 3b reports significant proportion of *NC* in **Bucket 1**. We observe similar results if we consider the author's increasing publication count (in place of h-index) on her copying behavior.

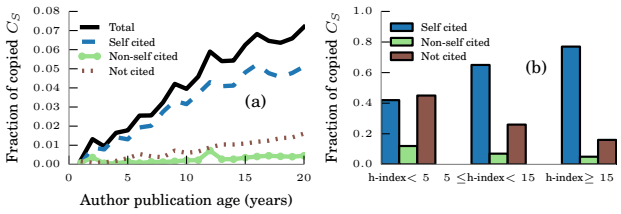

Figure 3: a) Average fraction of copied C_S over 20 years of author life span. Copying tendency increases as author matures. b) Proportion of three copying behaviors in three h-index buckets. Popular authors prefer to copy their own C_S.

5.3 Fieldwise Analysis

As described in section 2, our dataset consists of papers from 24 fields of computer science. In this section, we attempt to demonstrate the copying behavior in different fields of research. We randomly select 20,000 research papers from 24 distinct fields of computer science. The distribution of C_S in each field is shown in Figure 5a. Two interesting questions that require fieldwise analysis are:

[6] In order to keep the author experience same.

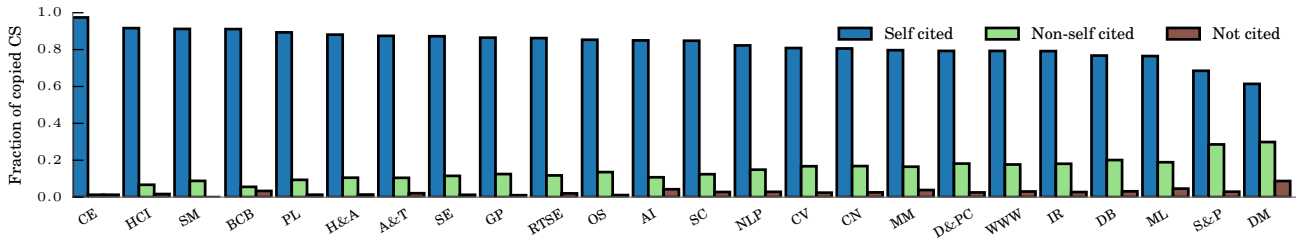

Figure 4: Proportions of three copying behaviors for 24 computer science fields. Majority of the copied contexts are originating from self citations. Applied fields have more tendency of copying and not citing than theoretical fields.

5.3.1 Is paper-wise copying behavior different in different computer science fields? Similar to the experiment in section 5.1.1, for each field, we divide copied C_S into three categories, *SC*, *NSC* and *NC*. Figure 4 presents proportion of three categories for 24 computer science fields. For all fields, majority of the copied contexts are originating from self citations. Next major proportion goes to non-self citations. Overall, we observe that applied fields have more tendency of copying and not citing than theoretical fields.

5.3.2 Is copying behavior same across all computer science fields? We perform this experiment along similar lines as the motivational study (see section 4), except that now papers are divided into 24 computer science fields[7]. Figure 5b presents cosine similarity distribution for two representative fields. We were surprised to find clear demarcation between theoretical fields[8] like, Algorithms & Theory etc., and applied fields like, Computer Networks etc. Even though, for small values of *CosSim* , all fields show similar behavior, for higher $CosSim(\geq 0.8)$ values (see inset Figure 5), theoretical fields show higher copying behavior as compared to applied fields. Note that, a sharp peak at $CosSim = 1$ is observed across all fields.

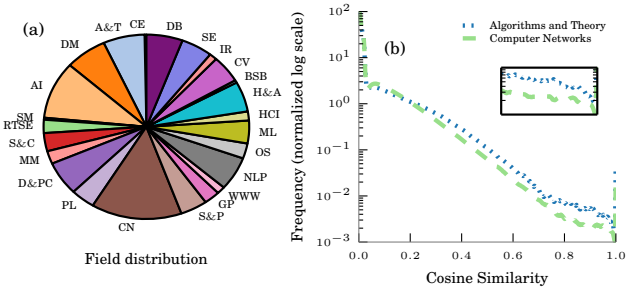

Figure 5: Fieldwise analysis: a) Distribution of C_S sampled from 24 fields. b) Comparison between cosine similarity distribution of two representative computer science fields. Inset shows significant difference between Algorithms & Theory (theoretical) and Computer Networks (applied) for higher $CosSim(\geq 0.8)$ values.

6 CONCLUSIONS

This paper has investigated micro-scale phenomenon of text reuse in scientific articles. Throughout the paper, we pose several interesting research questions and present an in-depth empirical analysis to answer them. We have provided further evidence that the theoritical fields indicate more copying than the applied fields. Finally, a number of potential limitations need to be considered. First, the current study employs a computer science dataset only. In the future, we plan to extend this study to other research fields as well. Second, cosine similarity metric may be too simplistic for this complex study. To further our research, we plan to employ word embeddings for more meaningful similarity computations. On account of the fact that the current work is only a preliminary attempt to understand micro-scale copying phenomenon of citation sentences, future extensions could possibly lead to rescaling of several popularity metrics such as h-index, impact factor etc. as number of times a paper has been cited may not be a true metric for impact of a paper or researchers in the research community. Alternatively, if this behavior is kept in check the quality of research can be expected to improve and the current popularity metrics will conform to the intuition behind their meaning.

[7] http://tinyurl.com/n2rwkbs
[8] https://en.wikipedia.org/wiki/Computer_science

REFERENCES

[1] Ioannis Arapakis, Yashar Moshfeghi, Hideo Joho, Reede Ren, David Hannah, and Joemon M Jose. 2009. Integrating facial expressions into user profiling for the improvement of a multimodal recommender system. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on.* IEEE, 1440–1443.
[2] Jonathan Bailey. 2014. University of Regina prof investigated for allegedly plagiarizing student's work. http://www.ithenticate.com/plagiarism-detection-blog/top-plagiarism-scandals-2014. (2014). [Online; accessed 24-January-2017].
[3] Jun-Peng Bao, Jun-Yi Shen, Xiao-Dong Liu, Hai-Yan Liu, and Xiao-Di Zhang. 2004. Semantic sequence kin: A method of document copy detection. In *Advances in Knowledge Discovery and Data Mining*. Springer, 529–538.
[4] Jun-Peng Bao, Jun-Yi Shen, Xiao-Dong Liu, and Qin-Bao Song. 2003. A survey on natural language text copy detection. *Journal of software* 14, 10 (2003), 1753–1760.
[5] Sergey Brin, James Davis, and Hector Garcia-Molina. 1995. Copy detection mechanisms for digital documents. In *ACM SIGMOD Record*, Vol. 24. ACM, 398–409.
[6] Tanmoy Chakraborty, Suhansanu Kumar, Pawan Goyal, Niloy Ganguly, and Animesh Mukherjee. 2014. Towards a Stratified Learning Approach to Predict Future Citation Counts. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '14)*. IEEE Press, 351–360.
[7] Daniel T Citron and Paul Ginsparg. 2015. Patterns of text reuse in a scientific corpus. *Proceedings of the National Academy of Sciences* 112, 1 (2015), 25–30.
[8] Eui-Hong Sam Han and George Karypis. 2005. Feature-based recommendation system. In *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 446–452.
[9] Guy Judge and others. 2008. Plagiarism: Bringing Economics and Education Together (With a Little Help from IT). *Computers in Higher Education Economics Review* 20, 1 (2008), 21–26.
[10] Michael Lesk. 2015. How many scientific papers are not original? *Proceedings of the National Academy of Sciences* 112, 1 (2015), 6–7.
[11] Noboru Nakanishi and Izumi Ojima. 1999. Notes on Unfair Papers by Mebarki et al. on"Quantum Nonsymmetric Gravity". *arXiv preprint hep-th/9912039* (1999).
[12] Dong-Chul Park and Young-June Woo. 2001. Weighted centroid neural network for edge preserving image compression. *IEEE transactions on neural networks* 12, 5 (2001), 1134–1146.
[13] Mayank Singh, Barnopriyo Barua, Priyank Palod, Manvi Garg, Sidhartha Satapathy, Samuel Bushi, Kumar Ayush, Krishna Sai Rohith, Tulasi Gamidi, Pawan Goyal, and Animesh Mukherjee. 2016. OCR++: A Robust Framework For Information Extraction from Scholarly Articles. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 3390–3400. http://aclweb.org/anthology/C16-1320
[14] Mayank Singh, Vikas Patidar, Suhansanu Kumar, Tanmoy Chakraborty, Animesh Mukherjee, and Pawan Goyal. 2015. The Role Of Citation Context In Predicting

Long-Term Citation Profiles: An Experimental Study Based On A Massive Bibliographic Text Dataset. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1271–1280.

[15] Matthew C Whitaker. 2005. *Race work: The rise of civil rights in the urban West.* U of Nebraska Press.

[16] Matthew C Whitaker. 2014. *Peace be Still: Modern Black America from World War II to Barack Obama.* U of Nebraska Press.

[17] Sven Meyer Zu Eissen and Benno Stein. 2006. Intrinsic plagiarism detection. In *Advances in Information Retrieval*. Springer, 565–569.