What Should We Do about Self-plagiarism

Miaomiao Zhang
University of Auckland
mzha029@ec.auckland.ac.nz
Student ID: 3186859

Abstract

In this paper, we will give the descriptions of what is self-plagiarism and what bad effects it has first. And then, we carefully examine ten pairs of published articles for evidence of self-plagiarism. Based on this evidence, we form a very rough estimate of the prevalence of various types of self-plagiarism in the computer science literature. We conclude our paper with a few opinions on the question of "what should we do about self-plagiarism?".

1. Introduction

It is well-known that plagiarism is a very serious problem, but how about self-plagiarism? Nowadays, self-plagiarism have received more public attention, because a lot of people realize that self-plagiarism is becoming more prevalent than other forms of scientific misconduct. At the same time, self-plagiarism also brings us a lot of bad effects on our research community. [1]

First, self-plagiarized papers occupy our limited publishable spaces.
Second, self-plagiarized papers are published again and again, while original paper cannot get the chance to be published, which will make people lose heart.
Third, self-plagiarized papers let our research community waste a lot of money on old results, rather than the new one.

Fourth, it is unfair that authors who self-plagiarize their papers get more prizes and academic credits, while authors who just publish their articles only once get less.
Fifth, a lot of near-identical papers make searching for information relevant to a particular topic harder than it has to be.
After we know that self-plagiarism has so many bad effects on our research community, generating some methods of what we should do about self-plagiarism becomes emergent. But before solving this problem, we should know what kinds of actions may be regarded as self- plagiarism and what is the most common type of self-plagiarism first.

2. What is self-plagiarism?

2.1 The first description of self-plagiarism

So what kinds of action should be regarded as self-plagiarism? There is one description said that self-plagiarism is using ones own previously published materials to create a new published material, but not crediting the previous paper as a source. This description just gives us a very rough expression of what is self-plagiarism. But it didn't mention the standard of how much authors use their original materials to create a new material belong to self-plagiarism. If author only uses one sentence from previously published work in his new article, does his new article belong to self-plagiarism? So somebody said that this description is not exact.

2.2 The second description of self-plagiarism

Comparing with the first description, the second one looks more clear and exact.
This description of self-plagiarism clarifies self-plagiarism into eight types: [1]
Textual reuse: Incorporating text/images/or other material from previously published work.
Semantic reuse: Incorporating ideas from previously published work.
Blatant reuse: Incorporating texts or ideas from previously published work in such a way that the two works are virtually indistinguishable.

Selective reuse: Incorporating bits and pieces from previously published work.
Incidental reuse: Incorporating texts or ideas not directly related to the new ideas presented in the paper.
Reuse by cryptomnesia: Incorporating texts or ideas from previously published work while unaware of the existence of that work.
Opaque reuse: Incorporating texts or ideas from previously published work without acknowledging the existence of that work.
Advocacy reuse: Incorporating texts or ideas from previously published work when writing to a community different from that in which the original work was published.
When it is believed that the actions are ethically or legally questionable, we replace reuse by plagiarism [1]. This description is much better than the first one, because the first one just gives us a very rough description, while this one gives us very detailed types of self-plagiarism, and let us know what kinds of action maybe belong to self-plagiarism.
But what is "ethically or legally questionable" and what is fair reuse, there is no obvious regulation.
In this paper, we will not talk about how to estimate articles belong to self-plagiarism or not, we just mention here, it is very difficult to give a definition of self-plagiarism and there is also no very clear border between fair reuse and self-plagiarism. Sometimes, self-plagiarism is just estimated by somebody's subjective opinion. All the decision will be made by editors. If editors think that this article is too similar with the last one published by the same author, they maybe regard it as self-plagiarism.
In the following, we will use the detailed types in the second description to analyze ten pairs of articles, and estimate the most common types of self-plagiarism roughly.

3. Most common types of self-plagiarism

In order to get the most common types of self-plagiarism, we choose ten pairs of articles (twenty articles) randomly and every pair of articles' similarity is more than 20%, mostly more than 35%. And then, we look into every pair of articles carefully and with statistic analysis, we roughly estimate the prevalence of various types of self-plagiarism in the computer science literature.

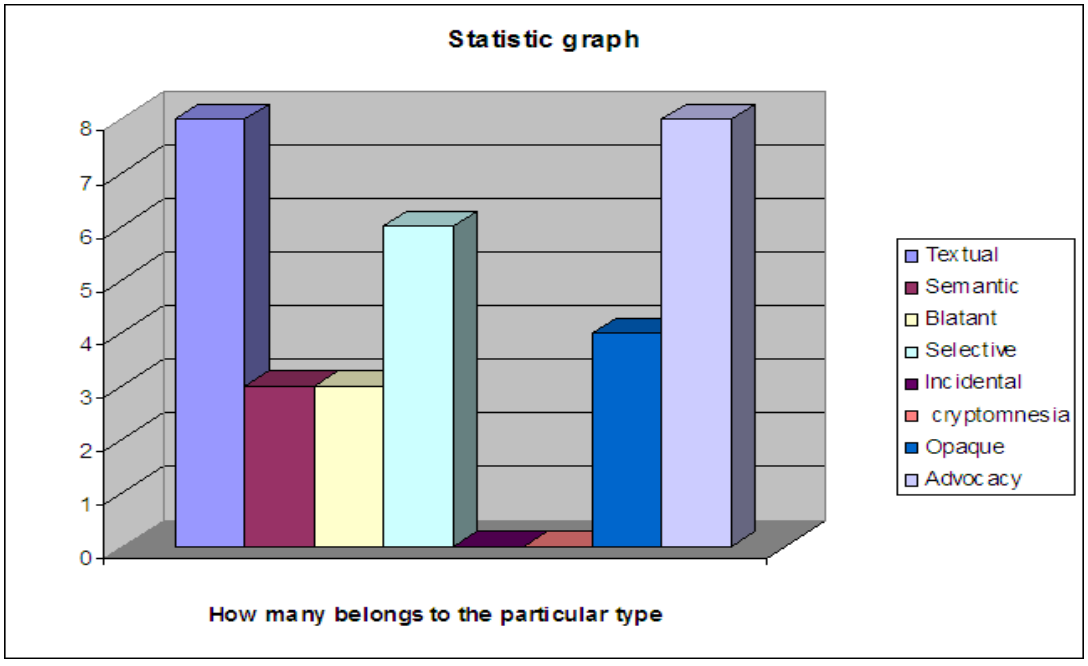
3.1 Statistic analysis of ten pairs of articles

In order to protect author's privacy, we will not mention titles of these articles and names of the authors. In the following, the table describes that in ten pairs of articles, how many pairs of articles use the particular type of self-plagiarism.

Textual	Semantic	Blatant	Selective	Incidental	Cryptomnesia	Opaque	Advocacy
8	3	3	6	0	0	4	8

From the table, we can see that eight pairs of articles use textual reuse and advocacy reuse. Selective reuse are used by six pairs of articles, and the number of article pairs which use semantic reuse is three, blatant reuse has the same number. One more pair of articles uses opaque reuse. No pair of articles uses incidental reuse and crytomnesia reuse. From the analysis of this table, we know that textual reuse and advocacy reuse are used by most of article pairs.

The graph in the following was drawn according to the statistic data in above table, which makes the comparison more clear and we can see the difference between them more easily.



From the graph, we also can get the same idea that textual reuse and advocacy reuse are used by the largest number of pairs. And the second one is selective reuse.

1	2	3	4	5
48.9%	20.1%	50%	36.6%	56.9%
Textual		Textual	Textual	Textual
		Semantic	Semantic	Semantic
		Blatant	Blatant	Blatant
Selective	Selective			
Opaque	Opaque	Opaque		
Advocacy		Advocacy	Advocacy	Advocacy
6	7	8	9	10
34.9%	25.9%	39.5%	31.4%	28.5%
Textual	Textual	Textual	Textual	
Selective		Selective	Selective	Selective
	Opaque			
Advocacy		Advocacy	Advocacy	Advocacy

And the above table displays what types of self-plagiarism are contained by each pair of articles and what is the similarity of each pair. We can see that each pair of articles not only contains one type of reuse, but several types at the same time. If one pair's similarity is high, this pair must use textual reuse and advocacy reuse. Among ten pairs of articles, only two pairs credit the previously published work as resource. That means most papers textually reuse the previously published paper and then publish them in different community, and there even exists one textual reuse paper really credits the previous paper as a resource. More than half number of pairs chooses selective reuse, which means the second paper just choose some relative parts or some research results from the previously published paper. Except textual reuse and advocacy reuse, selective reuse is also common.

3.2 Summary of above statistic analysis

In many ways, this study is not a very satisfactory study, because we just choose ten pairs of articles to analyse, not all articles can be chosen here, but we also can get the results what we want.

According to all analysis we discussed above, we cannot say that all pairs of articles are self-plagiarism, but we can make a rough estimate that textual reuse and advocacy reuse are the most popular type of reuse in similar articles. If the reuse can be replaced by self-plagiarism, we estimate that they should also be the most common types. The second one is selective reuse. In some pairs of articles, semantic reuse, blatant reuse and opaque reuse are also used, but they are not outstanding. That means, most authors of these articles use textual reuse, and then they post the second article to community which is different with the previously work published in.

So in the following, we will analyse how to let all people enforce high ethical standard and not self-plagiarize, how to prevent self-plagiarism, and how to detect textual reuse and advocacy reuse in self-plagiarized papers. Semantic reuse, blatant reuse and opaque reuse are not critical in this analysis, so we will not mention them in the following.

4. What should we do about self-plagiarism

In order to solve the problem of what we should do about self-plagiarism, in my opinion, we have two steps, one step is from subjective side, the other step is from objective side. From subjective side, we should correct people's concept and let all people realize that self-plagiarism is another kind of cheating, and enforce high ethical standard. If some of us still copy their own articles, from objective side, we should take some measures to prevent and detect.

4.1 From subjective side

In regard to ourselves:

1. We should know clearly what types of reuse constitutes self-plagiarism and what kind of reuse belongs to fair reuse before we write our assignments, reports or term papers.
2. We should know what the communities' policies about self-plagiarism are before we post our articles to publish (e.g. university's policy, ACM and IEEE's policies, etc.).
3. We should realize self-plagiarism is another type of cheating. Although plagiarism is more serious than self-plagiarism, self-plagiarism is also serious. In recent years, it has become a growing phenomenon and has detrimentally affected the entire learning community and research community.
4. In order to keep our own reputation, we should make sure we will not self-plagiarize our own published material. Even if we want to use some results of our own previously published paper in our new paper, we should credit the previously published work as a reference.
5. If we cooperate with other people to create a new article, we should carefully read other co-author's contributions to make sure that all authors of this article will not self-plagiarize themselves. You have the responsibility for other authors' contributions when you create one paper with someone else.

In regard to other people:

1. Inform other people about what kinds of action will be regarded as self-plagiarism and what kinds of action will be fair use.
2. Inform other people of all kinds of policies regarding self-plagiarism and always remind them.
3. Let other people realize how serious self-plagiarism is and how it detrimentally affects the entire learning community and research community.
4. If our own colleagues or classmates involved in self-plagiarism cases, we shouldn't relax our rules to deal with self-plagiarism problems. We should support the rules to enforce our ethical standards; otherwise, our self-plagiarism problem will become self-plagiarism crisis.

4.2 From objective side

Although we encourage all of us to avoid making mistakes about self-plagiarism, frequently, some people still involve in some self-plagiarism cases for their own purposes. We often hear of somebody publish the same result with minor modifications again and again. Sometimes, we found some papers and their contents are very close. So we can say that some people cannot reach our ethical standards, at this time, we should use strict policy to reduce the probability of self-plagiarism:

Currently, most universities have created their own policies about self-plagiarism. And different universities also have their own policies according to their own conditions. Also both ACM and IEEE have some policies about self-plagiarism.

ACM's policy is: "at least 25% of the paper is material not previously published; however, this is a somewhat subjective requirement that is left up to each publication to interpret". [1]

In November 2002, the IEEE Board of Directors approved a new policy on Duplicate Publication and Self-Plagiarism. This policy is found in the IEEE Policies document, Sections 6.4.1B(f) and 6.4.1B(h). These two sections are given below [5]:"
(f) Plagiarism is unacceptable. The verbatim copying or reuse of one's own research) as indicated in paragraph "h" below) is considered another form of plagiarism or self-plagiarism; it is unacceptable.
(h) Except as indicated in Section 6.3.4 (Multiple Publication of Original Technical Material in IEEE Periodicals), authors should only submit original work that has neither appeared elsewhere for publication, nor which is under review for another refereed publication. If authors have used their own previously published work(s) as a basis for a new submission, they are required to cite the previous work(s) and very briefly indicate how the new submission offers substantial novel contributions beyond those of the previously published work(s)."

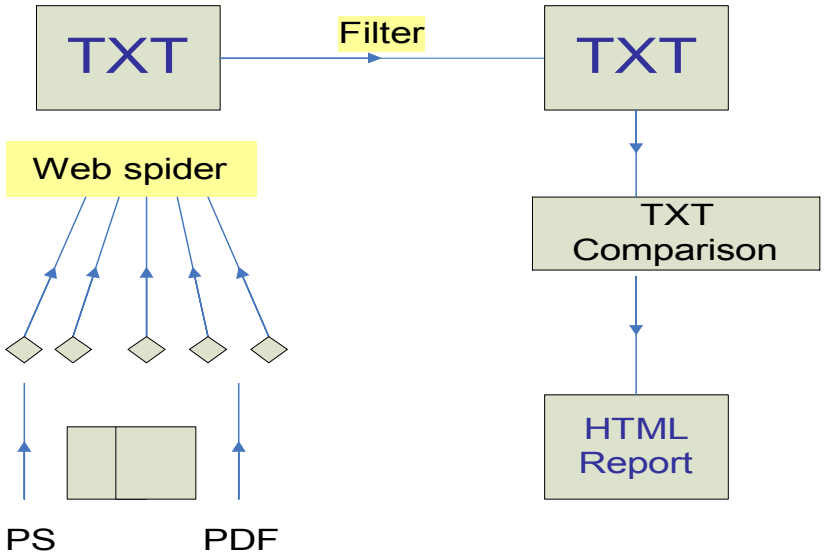
4.3 Using an automated system --- SPLaT to detect textual self-plagiarism. [2]

As we analyzed in section 3, we estimate that textual reuse and advocacy reuse are the most common types of self-plagiarism and selective reuse also another important aspect. So in order to know how many percents one paper self-plagiarizes the previously published one and whether they belong to self-plagiarism, we should detect them and get the similarity. In the following, let us talk about how to detect them.

a. Nowadays, we have a very popular self-plagiarism detection tool, whose name is SPLaT. It is convenient and easy to use.

The current version of SPlaT functions as a web spider that crawls through the web sites of fifty Computer Science departments, downloading research papers , after converting them to text, running a text analysis program to check for self-plagiarism and reporting pairs of papers with high textual overlap. This system is to search for instances of self-plagiarism by Computer Science academics.

Its working process will be described in the following graph:



According to the principal of this system, we can know that this system is good at detecting textual reuse. Because it needs to convert the PDF files into TXT and compare

them, different words have different TXT code, the copied parts always have the same TXT code. If we know that one article self-plagiarize the other, and these two articles are semantic self-plagiarism, when the system converts them to TXT code, it cannot recognize that they are self-plagiarism, because these two articles reveal the same idea but express in different way, they have different TXT codes after converting them. And the same idea can be used in selective reuse. Because selective reuse means that incorporating some parts and pieces from previously published work. That means, these copied parts or pieces appeared in two articles, so in SPLaT system, parts of their TXT codes are exactly the same. After using SPLaT to detect, we also can get similarity percentage in a HTML page and similar parts in these two articles will be colored in this page.

b. How to detect advocacy self-plagiarism

As we mentioned in section 2.2, advocacy reuse means incorporating texts or ideas from previously published work and then publish the second one to the community different from that in which the original work was published. In brief, if two very similar published articles are written by the same author and they are published in different community, which means advocacy reuse. Usually, if two works' materials are identical and the author post them to the same community, the editors can find that they are self-plagiarized works very easily, so it's nearly impossible for the author to publish them in the same community. Furthermore, in section 3, we can see that if two papers are self-plagiarism questionable, which means, the similarity of these two papers are very high, the author will post them to different communities. Then, in most of time, advocacy self-plagiarism belongs to almost all self-plagiarized works.

5. Conclusion of this paper

In this paper, we give two descriptions of self-plagiarism. According to the second description, we analyze ten pairs of articles and get some common types of self-plagiarism. And then, in some different ways, we talk about what we should do about self-plagiarism, how to prevent it and how to detect it.

Although all of us are encouraged to realize the detrimental effects of self-plagiarism for our research community, and shouldn't copy our own published papers to create a new one, we also can find some similar works published in different communities by the same author. That means, there still exist a lot of people who are self-plagiarize their own papers for some purposes.

So in order to prevent the happening of self-plagiarism in our research community, we have a lot of things to do in the future actually.

6. References

[1] Self-plagiarism in computer science Christian Collberg, Stephen Kobourov
1.April 2005 Communications of the ACM, Volume 48 Issue

[2] SPLAT: A system for self-plagiarism detection
http://splat.cs.arizona.edu/icwi_plag.pdf

[3] Join the fight against plagiarism
Mintzer, F.;
Signal Processing Magazine, IEEE
Volume 22, Issue 4, July 2005 Page(s):4 - 4
Digital Object Identifier 10.1109/MSP.2005.1458264

[4] Plagiarism, duplicate publication, and duplicate submission: They are all wrong!
Stone, W.R.; Antennas and Propagation Magazine, IEEE
Volume 45, Issue 4, Aug. 2003 Page(s):47 - 49
Digital Object Identifier 10.1109/MAP.2003.1241310

[5] Policy on Self-Plagiarism