Plagiarism Detection an Approach towards Uniqueness of Data: A Survey

[1]Prasad Dhekale, [1]Pradip Jamdade, [1]Swapnil Rahinj, [1]Shantanu Rai, [2]Shashikant Athawale

[1]BE Student, [2] Assistant Professor
Department of Computer Engineering,
AISSMS COE, Pune, India

***Abstract*: Plagiarism refers to intentionally or unintentionally stealing one's thoughts, theoretical work in particular field and claiming it as own. Thus, fraud writers gaining effortless advantage using the work of genuine writers without their consult. Plagiarism goes way back and has evolved with the fast use of internet. Various tools and softwares have been developed to identify such thefts. This paper puts light on the various types of the plagiarism, algorithms and the basic techniques they use to provide the similarity index of the documents, probable contents that get plagiarized. Also, the paper compares the famous plagiarism detection tools.**

***Keywords:* plagiarism, fraud writers, genuine writer, similarity index.**

I. INTRODUCTION

Plagiarism emerges to be one of the serious crimes which is committed by an individual when he/she steals or copies (as it is) the research contents own by someone else without their consent. These plagiarized contents can then be used to satisfy one's academical needs. However, plagiarism is not only limited to research and academics but can also be traced to multimedia contents like photos, videos, music also to blog posts and much more. According to research, about 70% of students confess that they used the plagiarized content in their academic work thus having a negative impact on learning process [1]. Attempts should be made in this field to eliminate such practice so that each individual work is genuine. Hence various plagiarism detection tools and software have been developed to help reduce such piracy practice to a great extent and determine the uniqueness of the data. The impact of such software depends on the type of algorithms used by the software for detecting the plagiarized content.

Rest of the paper is organized in following format. Section 2 provides the information about the fields in which plagiarism is common practice. Section 3 discusses about the types of the plagiarism. Section 4 discusses about the various algorithm used. Section 5 lists various plagiarism softwares with their advantages and disadvantages.

II. AREAS OF IMPACT

Although the plagiarism is commonly detected in academic works, it is not limited only to it.

- **Plagiarism on Academy**

This is a most common area of plagiarism. Students extensively use the readymade contents easily available for their academic works. Such content may have copyright and offenders may be punished for the act of piracy. Plagiarism may also happen accidentally when failing to mention proper citations, but still is treated as plagiarism.

- **Plagiarism on World of Internet**

With extensive use of internet to easily access contents online, plagiarism is growing rapidly. Website owners or bloggers tend to copy lots of data from other websites and publish it on their own website or blog.

- **Plagiarism on Research**

Plagiarism is also seen in research where researchers steal other researcher's thoughts or their work content who is working in the same field. Researchers may also copy data from multiple sources and then combinedly represent it as their own.

- **Plagiarism on the Arts**

Multimedia contents such as music, videos and pictures also when copied are termed as plagiarized. When submitting or using a picture without the owner's consent is also plagiarism. Similar is for videos and music. [2]

Work in any of the above areas depict the efforts taken by an individual or group. Plagiarism in any of the above areas may lead the victim to a serious penalty.

2.1 Content getting plagiarized

Based on the areas of plagiarism the common content getting plagiarized are as follows:

1. Research Papers/documents.
2. code/code snippets.
3. Images/Multimedia Files.
4. Algorithms. [3]

III. TYPES OF PLAGIARISM

There are various types of plagiarisms which can be classified based on the gravity or frequency plagiarism, amount of content getting plagiarized and also based on whose content is being plagiarized(self/other). Firstly, there is an "direct plagiarism", which represents exact replica of a content done by an individual, that is copying the exact content word by word and stating as own [4]. "complete plagiarism", here the entire document of other individual is presented as self-work. "Paraphrasing plagiarism" which involves making some minor changes in the writings of already published documents of other publishers [5]. Although some part of the document is self-written, but the language used is of the original author's only with some small changes and hence the content is still plagiarized. All the previously mentioned types of plagiarism fall under the category of amounts of content being copied. [5] "Self-plagiarism" another type of plagiarism which is experienced when a research individual or student reuses his or her previously published content or part of the content without explicit reference. Many institutions have certain criteria about the amount of previously published content that can be reused. Also, there is "accidental plagiarism" which happens when the person does not include proper citations, the sources from where the content is taken or include some paraphrasing. Accidental plagiarism is taken with equal degree of seriousness as that of the actual plagiarism. [6]

IV. ALGORITHMS AND TECHNIQUES USED

There are couple of algorithms that are used for plagiarism and this section will mainly focus on the main techniques used by these algorithms. Commonly at the core all the algorithm uses some sort of pattern matching.
The general algorithm for Plagiarism detection processes the document or set of strings in following steps:

1. Remove all the words or letters not important in pattern matching e.g. to, the or comments in case of code etc.
2. Convert the remainder words to their root meaning.
3. Perform string matching using algorithms like parallel KMP.
4. Calculate the percent match and store the results in a file e.g. .xml
5. Compare another document.

4.1 Stemming and KMP

Stemming Algorithm can be used to convert words or sentences into their root or more general meaning. [7] Using this technique paraphrasing type of plagiarism can also be detected which adds more to the accuracy of the system. An algorithm may use Knuth-Morris-Pratt (KMP) string matching algorithm to match a set of string S1 in one document to another string S2 in another document. The algorithm is based on forward pattern matching [8]. This algorithm pattern matches each string of S1 to each string of S2 thus providing accurate match but requiring more time to provide such results. Hence a parallel KMP can be used to process more number strings at a time. [9] Aho-Corasick is another pattern matching algorithm that perform a bulk string matching. Both the Aho-Corasick and KMP are suitable for both monolingual and cross-lingual plagiarism detection [10]. Fingerprinting is another technique used for identifying plagiarized content.

4.2 Fingerprinting

In fingerprinting [11] set of the substring from the document is selected and we have to calculate the index of the document this index also called as minutiae. It is a set of integers that act as fingerprint for the document. Then this document is compared with relevant document with preprocessed index. According to the index comparison between document then we can understand that whether that document is plagiarized or not. A hash function is used to generate the index for the document i.e. an minutiae. [11] This hash function should provide results within acceptable time and also the resultant minutiae should be reproducible. Following hash function satisfies this criterion:
For each word or string, $s_0,s_1...s_n$ and a function for converting character to string i.e. $Z(s_n)$ then for each character s_i in the string, we can compute,

$$h(s_i) = h(s_{i-1}) \oplus (Z(s_i) + h(s_{i-1}) << 6 + h(s_{i-1}) >> 2),$$

and if we get max as M, the hash value we get is, $h(string) = h(s_k) \bmod M$. [11]

4.3 Levenshtein distance

The minimum distance requires to convert one string into another string is called as Levenshtein distance [12]. In Levenshtein distance algorithm we may have to convert the words into integer i.e. We have to assign integer value to each word and from comparison between this integer values we can easily detect the plagiarism. This is a effective string approximation tool. [13]
According to this survey most of the plagiarism detection algorithms or tools make use of APIs from well-established search engines like Google, Bing or Yahoo for efficient pattern matching without compromising much of the accuracy. But this technique may not search the actually required database that may have the right content to be searched.

V. COMPARISON OF DIFFERENT TOOLS AVAILABLE ONLINE

Table 1. Comparison Table of Different Plagiarism Tools Available Online

Software	Search Method	Pros	Cons	Number of words (free version)
1.Turnitin	-This Software has Own Database named as "Turnitin" Which include information from all the fields. -Search occurs in Turnitin as well as Internet.	- It makes relatively very easy identifying the plagiarism - Provide different facilities like Precheck, Grade Check. - Provide Guidelines regarding the usability about the software.	- failed to check different citation styles like APA, MLA. - It also Cannot check the figures, tables are plagiarized or not.	-It gives only one credits to one use, but it has 2500-word limit.
2.iThenticate	-This Software searches over the Internet, as it does not have special database.	- This Software also provide deep internet checking including live and cached links. [16] - It provides plagiarism checking in books, documents, magazines and newspapers.	- Synonym and Syntax checking are not supported. [16] - Does not support to the previously uploaded data. [14] [15] [16]	- It does not give any free trial to the user.
3.Dupli Checker	-It Searches on the Web-based data. [17]	- It provides cost free service. - It is very easy to use via two step(copy-paste) mechanism.	- Limited number of searches is permitted. - Not suitable for long document search due to small word search limit.	-It gives 1000 words per search.
4.Copyleaks	-Mainly Focuses on the Cloud based Authentication. -Makes use of internet for efficient search. [18]	- Easy to identify theft by using URL on Copyleaks. - Support different types of file format such as txt, pdf, docx etc. - Short paragraphs are free cost searched. - Time Efficient	- Copy paste not possible as it works on the URL of the document. - without registration it is unable to use. -credit card details are required.	-It gives 10 pages for free per user every month after sign in.

5.Plagiarism Checker	-Plagiarism Checker software searches documenting the internet to find out documents plagiarized or not. [19] [20]	-Easy to use - Provide standard guidelines for the user. - Mostly prefer due to simplicity.	- Supports only for Google or Yahoo browser. - Unable to find plagiarized image. - Grammar checking is not performed. [20]	-It gives 1000 word per user limit.

V. CONCLUSION

In this paper we have made an attempt perform survey on different algorithms and tools available to detect plagiarism. The paper also discusses various types of plagiarism such as direct plagiarism, paraphrasing plagiarism, complete plagiarism, self-plagiarism and accidental plagiarism. Some pattern matching algorithms such as KMP, Levenshtein distance used in detection techniques such as Data source comparison and Manual search for strings. Also, survey concludes that 100% plagiarism detection is hardly possible.

REFERENCES

[1] Amruta Patil, "Survey on different plagiarism detection tools and software": International Journal of Computer Science and Information Technologies, Vol. 7 (5) , June 2016.
[2] https://www.wisconsinhealthconnection.com/how-does-plagiarism-occur-in-different-areas/
[3] A. S. Bin-Habtoor and M. A. Zaher, "A Survey on Plagiarism Detection Systems", International Journal of Computer Theory and Engineering, Vol. 4, No. 2, April 2012.
[4] Dharmesh Namdev and Jayesh Surana, "Plagiarism Detection techniques", International Journal of Computer Applications (0975 – 8887) International Conference on ICT for Healthcare (ICTHC 2015).
[5] https://www.bowdoin.edu/studentaffairs/academic-honesty/common-types.shtml.
[6] https://www.enago.com/academy/fraud-research-many-types-plagiarism/.
[7] https://pdfs.semanticscholar.org/1c0c/0fa35d4ff8a2f925eb955e48d655494bd167.pdf
[8] R Yu Tsarev IOP Conf. Ser.: Mater. Sci. Eng. 122 012034, "KMP and Boyer-moore algorithms" 2016.
[9] Panwei Cao and Suping Wu, "Parallel research on KMP algorithm" 2011.
[10] S. M. Alzahrani, N. Salim, A. Abraham, Understanding plagiarism linguistic patterns, textual features, and detection methods,IEEE Transactions On Systems, Man, and Cybernetics, Part C(Applications and Reviews)42(2)(2012).
[11] Hoad T. C., Zobel J. " Methods for Identifying Versioned and Plagiarized Documents". JASIST 54(3): (2003)
[12] Zhan Su, Byung-Ryul Ahn, Ki-yol Eom, Min-koo Kang, Jin-Pyung Kim and Moon-Kyun Kim,"Plagiarism Detection Using the Levenshtein Distance and Smith-Waterman Algorithm" : June 18 - 20, 2008
[13] Rishin Haldar and Debajyoti Mukhopadhyay, "Levenshtein Distance Technique in Dictionary Lookup":2-5 January 2011.
[14] https://www.turnitin.com/blog/top-15-misconceptions-about-turnitin.
[15] http://www.checkforplagiarism.net
[16] Asim M. El Tahir Ali, Hussam M. Dahwa Abdulla, and V´aclav Sn´aˇsel, "Overview and Comparison of Plagiarism Detection Tools".: January 2011.
[17] http://www.duplichecker.com/, Duplichecker
[18] https://copyleaks.com/education/dashboard
[19] http://www.plagiarismchecker.net/plagiarism-checker-reviews.php
[20] http://plagiarism-checker-review.toptenreviews.com/index.html