International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

PLAGIARISM CHECKER USING - NLTK

Allu Sravani (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.
Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

Plagiarism is when someone takes another author's works, thoughts, ideas, etc. without proper referencingand claim it as his/her own works. Plagiarism detection is the process to find the plagiarism within a work or documents. Withthe advance of modern technology, it makes it easier for people to search for information and plagiarize the work of others. Although the effort and ideas for an image-based plagiarism detection has been increasing over the years, flaws are still presence in the current systems. This paper proposes a new system that can cover those flaws. It consists three stages: the pre-processing, feature extraction and comparison stage. The results showed in an ascending order of similarity index and true and false. However, the accuracy is 100% in case of unedited images and variated in other operations such as flipped, rotated, greyscales and cropped

Keywords :Image Plagiarism, Image Retrieval, Feature extraction.

1.INTRODUCTION

Plagiarism basically means the wrongful stealing of an author's work, thoughts, ideas, etc. and claiming it as your own original work. Plagiarism is considered as deceit and a breach of ethics. In academics, students that are caught with plagiarism are exposed to various level of penalties and punishment and may even lead to expulsion. Plagiarism in itself cannot be considered as a crime but as copyright violation. In the academics and other industries that are sensitive to copyright infringement, plagiarism is grave misconduct in integrity. The law cannot and usually will not punish plagiarism, but it is up to the institution on how to handle it once it happens [1]. Plagiarism detection is usually split into two which is text-based plagiarism detection and
image-based plagiarism detection. For text-based plagiarism detection there are currently five techniques that is used most often in different fields. These techniques are Fingerprinting,
String Matching, Bag of Words, Citation Analysis and Stylometry. String Matching is mostly used in computer science where it compares the documents words for words. Bag of words represents the documents in one or two vectors for comparison. Citation analysis is mainly used in scientific texts because it only compares the citation and reference of the documents. Stylometry check the author's unique writing style for detection author's ownership [2]. For image-based plagiarism detection, there is no commonly used techniques like the text-based plagiarism detection, but they usually share the same processes and steps. When we say plagiarism checking or detection we usually mean checking only the text in the file or document for plagiarism. Most of the times when you check your documents or files for plagiarism through a plagiarism

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

checker software they will check for images and then discard them. This is one of the fatal flaws that the current system is facing. In the field of research, images and flowchart can carry vital piece of information that can easily be plagiarized if the flaw in the system is there.

2. PLAGIARISM

From the Oxford Dictionary, plagiarism means the act of taking someone else's work, ideas, thought, etc. andclaiming it as your own work. The word plagiarism comes from the Latin word Plagiarus which means kidnapper, plunderer or seducer. The word Plagium which means
kidnapper is derived from the word Plaga which means to capture or trap. Modern days the word plagiarism means to plagiarize. The process of checking a work or documents for plagiarism is called Plagiarism Detection or Checker [3]. The history of plagiarism first began from religious
texts where most of them were authorless, so it is copied extensively and merged into later works. At the mid-1600, it is very common for there to be accusation of plagiarism for every creative field [3]. In the year 1709, the first copyright law was passed but it has more to do with protecting the publisher's right than the author's, but another law was passed soon after that to protect author's right. James Boswell, who is also known as the biographer for Samuel Johnson, was a lawyer that opposed how long the copyright of the author lasted which at that time ended up to 21 years [4]. In the beginning of the 19th century, the laws for copyright is pretty like what we have today. The only difference is the issue of enforcing those laws across the borders. Most European country sign an agreement to prevent book piracy except for America which signs it at the year 1891 [4].

3. PLAGIARISM DETECTION TECHNIQUE

A.Text-Based Plagiarism Detection Technique

The main technique used for text-based plagiarism detection is, Fingerprinting, String Matching, Bag of Words, Citation Analysis and Stylometry. The most used technique is the Fingerprinting technique where the system will select a set of multiple substrings from the documents and the sets signifies the fingerprints which is made up of the elements called minutae. The plagiarism checking is done by taking the fingerprints of the documents and comparing them to a String matching is mainly used in the computer science field where the system will compare word for word on each document. This system detect plagiarism in a pair with the original documents and with the collection of
references. Although plentiful methods have been proposed but this system is still computationally expensive making it unfit for large number of documents. For Bag of Word
technique, it used a vector space retrieval representation where the documents are represented with one or two vectors to be used for similarity comparison. The system can use the regular cosine similarity measure or further advance similarity comparison technique. Citation analysis more widely

used for checking plagiarism in scientific text because it is the only technique that does not use textual analysis but examines the citation and references of the documents to recognize comparable pattern. This technique is still considerably new, so it is not ready for commercial used yet[2]. The last technique, Stylometry detects plagiarism by checking the writer's unique writing style so it is more widely used for checking the original owner attribute. The system compares the stylometric models for different text segment that are stylistically different from others.

B. Image-Based Plagiarism Detection Technique

Image-based plagiarism detection is less used compared to text-based plagiarism detection, so it does not have a widely popular technique that is used everywhere. Instead, they are still researching for a good method to detect plagiarism in images. Here we will discuss a few of the
methods that was proposed every year. Method by Popescu and Farid [5] proposed a system by using the Principal Part Analysis (PCA). This system divides the numerous tiny sized blocks into vectors which is then organized lexicographically before matching them. The main drawback for this system is that if the image quality is too low then the accuracy will fall as well. Mahdian and Saic [6] apply a blur movement variant to signify the image region so that the images will not bedegraded from blurring and noises. This system begins by tilting the image with selected size blocks and defined it with blur invariants. The drawback from this method is that the computation time for this system is very long. An experiment conducted by Wang et al [7] used a
copy-move plagiarism detection system by applying the victimization Hu moments to cut down the computation time of the system. This system divides the image into numerous sized blocks and then applying the Hu moments on the block they computed the Eigen value. A more enhanced method has been proposed by Zimba and Xingming [8] for the copy-move detection system. The system starts by converting the image into a gray scale image and then applying DWT. The image is then divided into overlapping blocks and then PCA is done to each block. This method cuts down the computation time of the system compare to the PCA method by reducing the size of the image. Bravo-Solorio and Nandi [9] proposed a system to detect reflection, scaling and rotation of an images. The drawback is that their methods produced a lot of matches which needed to be further improve. A system that had been proposed by Sridevi et al [10] used the copy-move detection system in parallel which makes the system unable to used methods that has a requirement of long computation time. The only disadvantage with this system is that it cannot detect colored images.

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

4. RELATED WORK

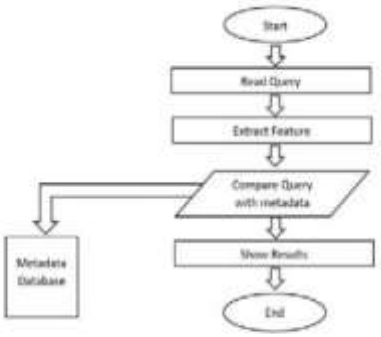

Figure 1: Flowchart of shape-based flowchart detection [11]

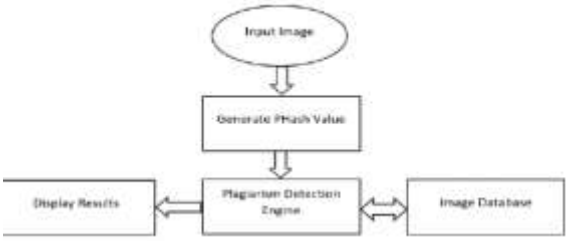

Figure 3: Flowchart of Perceptual Hash [13]

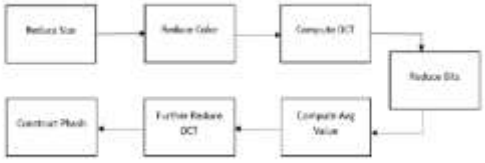

Figure 4: Workflow of Phash Generation [13]

5. CONCLUSION

Plagiarism detection is a well-known phenomenon in the academic arena. Copying other people is considered as a serious offence that needs to be checked. In this paper, an enhanced system to detect the Plagiarism of Images was proposed. The feature extracted from images and saving it to
the databases is color using the RGB and HSV color space, texture using Tamura texture and then shape using canny edge algorithm. The results showed in an ascending order of similarity index and true and false. However, the accuracy is 100% in case of unedited images and variated in other operations such as flipped, rotated, greyscales and cropped.

6. REFERENCES

[1]. Green, Stuart P, Plagiarism, Norms, and the Limits of Theft Law: Some Observations on the Use of Criminal Sanctions in Enforcing Intellectual Property Rights, Hastings Law Journal, 2002, 54 (1)

[2]. How Plagiarism Detection Works. (2016, May 18). Retrieved April 17, 2018, from https://www.plagiarismtoday.com/2016/05/03/plagiarismdetection- works/

[3]. Vinod K.R.*, Sandhya.S, Sathish Kumar D, Harani A, David Banji and Otilia JF Banji, Plagiarism – History, Prevention and Detection, journal for drugs and medicines, 2011, 3(1), 1-4

[4]. Eshgh, A. (n.d.). Copyright Timeline: A History of Copyright in the United States. Retrieved April 17, 2018, from http://www.arl.org/focus-areas/copyright-ip/2486-copyright-timeline

[5]. AC Popescu and H Farid, Exposing Digital Forgeries by Detecting Duplicated Image Regions, Dept Computer Science, Dartmouth College, Hanover, 2004, 515.

[6]. B Mahdian and S Saic, Detection of Copy–Move Plagiarism using a Method based on Blur Moment Invariants, Forensic Science International, 2007, 171(2), 180-189.

[7]. JW Wang, GJ Liu, Z Zhang, Y Dai and Z Wang, Fast and robust forensics for image region-duplication Plagiarism, Acta

Automatica Sinica, 2009, 35(12), 1488-1495.

[8]. M Zimba, and S Xingming, DWT-PCA (EVD) Based Copy-move Image Plagiarism Detection, International Journal of Digital Content Technology and its Applications, 2011, 5(1), 251-258.

[9]. S Bravo Solorio and AK Nandi, Automated Detection and Localisation of Duplicated Regions Affected by Reflection, Rotation and Scaling, Image Forensics Signal Processing, 2011, 91(8), 1759-1770.

[10]. M Sridevi, C Mala and S Sandeep, Copy–Move Image Plagiarism Detection, Journal of Computer Science and Information Technology, 2012, 52, 19-29.

[11]. Senosy Arrish, Fadhil Noer Afif, Ahmadu Maidorawa and Naomie Salim, Shape-Based Plagiarism Detection for Flowchart Figures in Texts, International Journal of Computer Science & Information Technology (IJCSIT), 2014, 6(1).

[12]. Prajakta Ovhal, Detecting Plagiarism in Images, International Conference on Information Processing (ICIP), 2015.

[13]. Vipul Bajaj, Sanket Keluskar, Ravi Jaisawal and Prof. Rupali Sawant, Plagiarism Detection of Images, International Journal of Innovative and Emerging Research in Engineering, 2015, 2(2).

[14]. Siddharth Srivastava, Prerana Mukherjee and Brejesh Lall, imPlag: Detecting Image Plagiarism Using Hierarchical Near Duplicate Retrieval, IEEE INDICON, 2015.

[15]. Jithin S Kuruvila, Midhun Lal V L, Rejin Roy, Tomin Baby, Sangeetha Jamal, Sherly K K*,

[16]. Flowchart Plagiarism Detection System: An Image ProcessingApproach, International Conference on Advances in Computing & Communications, 2017.