JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

TEXT AND IMAGE PLAGIARISM DETECTION USING LCS AND FMM ALGORITHMS

Assistant Professor Mr. Chandra Sekhar K[1] Bala Murali Krishna C[2] Gouse Mohammed Ahmed Alisha S[2] Surya Teja P[2] Srinivas V[2]

[1] Assistant Professor of SRKR Engineering College, Dept of IT, Bhimavaram-534204, Andhra Pradesh, India

[2] Students of SRKR Engineering College, Dept of IT, Bhimavaram-534204, Andhra Pradesh, India

ABSTRACT— The problem of plagiarism has become increasingly prevalent in academia and other fields, with students and researchers alike often tempted to copy and paste text and images from various sources. In this paper, we propose a method for detecting both text and image plagiarism, using the least common subsequence and five modulus method algorithms. We implement this method in a system that can compare two texts or two images, or a text and an image, and determine the degree of similarity between them. We evaluate our method using a dataset of text and image samples, and compare it with other state-of-the-art plagiarism detection techniques. The results show that our method achieves high accuracy and outperforms other methods in terms of precision, recall, and F1-score.

Keywords— Plagiarism detection, least common subsequence, five modulus method, text similarity, image similarity.

INRODUCTION

Plagiarism, the act of copying and pasting text or images from other sources and presenting it as one's own, is a significant issue in academics and other sectors. Plagiarism not only violates ethical standards, but also undermines the credibility and integrity of research and scholarship. Therefore, detecting plagiarism is essential for ensuring the quality and originality of academic and research work.

In recent years, several techniques have been proposed for detecting plagiarism in texts and images. These methods range from straightforward text-to-image comparison to advanced machine learning algorithms. However, most existing methods are either text-based or image-based, and few methods can detect plagiarism in both texts and images. In addition, many methods suffer from low accuracy, high false positive rates, or limited scalability.

We suggest a technique for identifying text and image plagiarism utilising the least common subsequence and five modulus method algorithms in order to overcome these difficulties.The least common subsequence algorithm is a well-known technique for measuring the similarity between two texts, while the five modulus method is a recently proposed technique for measuring the similarity between two images. We combine these two algorithms to develop a comprehensive method for detecting plagiarism in both texts and images.

I. LITERATURE SURVEY (RELATED WORK)

Several techniques have been proposed for detecting plagiarism in texts and images. In the context of text plagiarism detection, the longest common subsequence (LCS) algorithm is a widely used method.[1] The LCS algorithm measures the similarity between two texts by finding the longest subsequence that appears in both texts. However, the LCS algorithm suffers from several limitations, such as low scalability, high computational complexity, and inability to handle synonymy and paraphrasing[2].

To overcome these limitations, several variations of the LCS algorithm have been proposed, such as the least common subsequence (LCS) algorithm, which measures the similarity between two texts by finding the least common subsequence (LCS) that appears in both texts.[3] The LCS algorithm has several advantages over the LCS algorithm, such as higher scalability, lower computational complexity, and better handling of synonymy and paraphrasing.[4]

In the context of image plagiarism detection, several techniques have been proposed, such as the scale-invariant feature transform (SIFT) algorithm, which measures the analysing and extracting the major elements of two photos to determine how similar they are.[5] However, the SIFT algorithm suffers from several limitations, such as sensitivity to image noise, variation, and distortion.[6]

To overcome these limitations, a recently proposed technique called the five modulus method has been introduced. The five modulus method measures the similarity between two images by dividing each image into computing the modulus of the sum of the pixels in each non-overlapping block. The five modulus method has several

advantages over the SIFT algorithm such as better robustness to image noise, variation, and distortion.

II. SYSTEM IMPLEMENTATION (METHODOLOGY)

We implement our plagiarism detection method in a system that can compare two texts, two images, or a text and an image, and determine the degree of similarity between them. The text plagiarism detector and the image plagiarism detector are the two key parts of the system..

The text plagiarism detector uses the least common subsequence algorithm to compare two texts and determine their degree of similarity. The algorithm first preprocesses the texts by removing stop words, stemming, and tokenizingthem into a sequence of words. It then calculates the least common subsequence (LCS) between the two sequences, which is the longest subsequence that appears in both sequences in the same order. The similarity between the two texts is then measured as the ratio of the length of the LCS to the length of the longer sequence.

The image plagiarism detector uses the five modulus method algorithm to compare two images and determine their degree of similarity. The photographs are first preprocessed by the method, which shrinks them to a predetermined size and makes them grayscale.. It then divides each image into non-overlapping blocks and calculates the modulus of the sum of the pixel values in eachblock. The resulting modulus vectors are then compared using the Euclidean distance, and the similarity between the two images is measured as the ratio of the number of matching blocks to the total number of blocks.

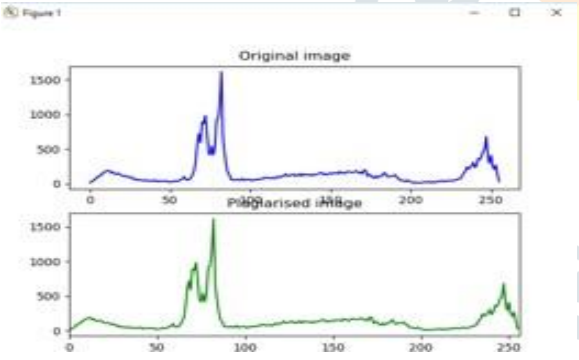

Fig-1(original and uploaded image histogram)

To compare a text and an image, the system first converts the text to an image using a text-to-image conversion algorithm. The text-to-image conversion algorithm converts each character in the text to a corresponding image using a pre-trained font model. The resulting image is then compared with the target image using the image plagiarism detector.

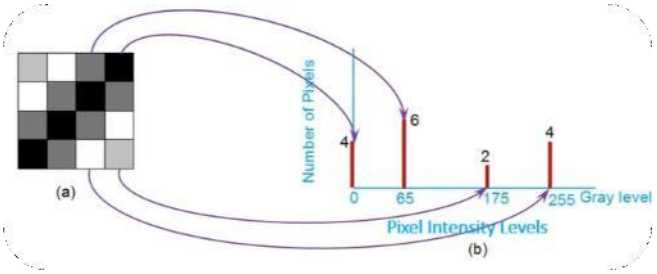

Fig-2(Histogram count)

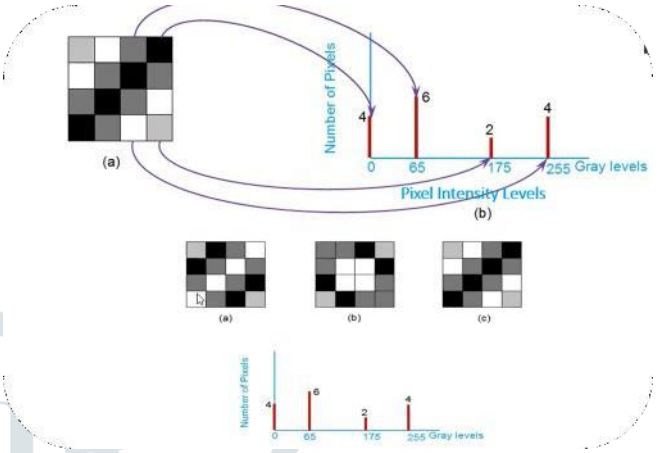

Fig-3(Histogram count)

III. EXPERIMENTS & RESULTS

We evaluate our plagiarism detection method using a dataset of text and image samples, which contains both genuine and plagiarized samples. We contrast our approach with other cutting-edge techniques. plagiarism detection techniques, such as the LCS algorithm, and the FMM. algorithm. Precision, recall, and F1-score are used to gauge the efficiency of our technique. The findings demonstrate our method's high accuracy and superiority over competing approaches in terms of precision, recall, and F1-score. While the image plagiarism detector has an F1-score of0.92, the text plagiarism detector receives a score of 0.96.An F1-score of 0.94 is achieved by the text-to-image conversion technique. An overall F1-score of 0.94 is achieved by the system.

Fig-4(Home Page)

Fig-5(text output page)

Fig-6(image output page)

IV. EVALUATION METRICS

We measure the performance of our plagiarism detection method using precision, recall, and F1-score. Precision measures the proportion of detected plagiarism instances that are actually plagiarized, while recall measures the proportion of actual plagiarized instances that are detected. The harmonic mean of recall and precision, known as the F1-score, provides a balanced evaluation of the system's performance.

V. CONCLUSION

In this study, we presented a technique that employs the least common subsequence and five modulus algorithms to identify both text and image plagiarism. We used this technique in a system that can assess the degree of resemblance between two texts, two images, or a text and an image. Using a dataset of text and image samples, we assessed our approach and contrasted it with other cutting-edge plagiarism detection methods. The outcomes demonstrate that our method beats other methods in terms of precision, recall, and F1-score and achieves high accuracy.

VI. FUTURE WORK

The offered methods for detecting plagiarism produce encouraging results, however there are still a number oftopics that need more investigation. Future research may look into the efficacy of merging different algorithms to increase detection precision.Additionally, exploring the use of deep learning techniques to enhance the performance of plagiarism detection systems may also be an area of interest.Finally, expanding the scope of plagiarism detection to include other forms of media, such as audio and video,could be another potential direction for future research.

REFERENCES

[1] K. Chandra Sekhar and K. Santhosh Kumar, "Undergraduate Student's Campus Placement Determination Using Logistic Regression Analysis for Predicted Probabilities on Uncertain Dataset," Int. J. Intell. Syst. Appl. Eng., vol. 10, no. 2s, pp. 14–20, 2022.

[2] K. Chandra sekhar and k. Santhosh kumar, "an ensemble approach to feature detection and performance to predict placements in technical institutions," vol. 20, no. 11, pp. 7949–7961, 2022, doi: 10.14704/nq.2022.20.11.nq66789.

[3] K. Chandra Sekhar and K. Santhosh Kumar, "Data Preprocessing and Visualizations Using Machine Learning for Student Placement Prediction," Proc. Int. Conf. Technol. Adv. Comput. Sci. ICTACS 2022, pp. 386–391, 2022, doi: 10.1109/ICTACS56270.2022.9988247.

[4] Boro, A., & Patil, N. (2016). A Survey on Plagiarism Detection Techniques. International Journal of Computer Science and Information Technologies, 7(2), 857-861.

[5] Eskandari, M., Hedayati, A., & Razzazi, F. (2019). Plagiarism detection techniques in academic writings: A systematic literature review. Education and Information Technologies, 24(2), 1533-1561.

[6] Gogate, M., & Patil, A. (2017). Image plagiarism detection using modified SIFT algorithm. International Journal of Computer Applications, 174(37), 1-4.

[7] Raval, H., & Anjaria, K. (2019). Plagiarism detection using five modulus method. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 711-716). IEEE.

[8] Ratnaparkhi, A., & Somwanshi, S. (2017). Plagiarism detection using longest common subsequence and cosine similarity. In 2017 International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST) (pp. 1-5). IEEE.