INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

DETECTION OF PLAGIARISM USING ARTIFICIAL NEURAL NETWORKS

V SRIKANTH
MCA, MTECH, MBA AND PGDBM

ABSTRACT— Today, much more than in the past are discussed of plagiarism in the research. Conditions of the Web and Possibility of complex and smart searches ina short time, are rated to this, and as a result has arrived significant damages to the research. Tools designed to deal with plagiarism act on the text and ignore images. On the other, an inseparable part of information transfer is images that transfer the large volume of information in an article or scientific research. Because of the imagesinclude a very wide range and especially found large amounts of flowchart images in the computer's texts, and as respects, flowcharts are carrying a lot of information, could be one of the options of plagiarism. The purpose of this paper is examining the plagiarism rate of a paper in terms of flowchart images plagiarism using artificial neural network. The average of flowchart images recognition accuracy in terms of structure, nodes and edges in the proposed method with 81.91 percent, indicating thesuccess of this method.

1. INTRODUCTION

Academic plagiarism has been defined as –the use of ideas, concepts, words, or structures without appropriately acknowledging the source to benefit in a setting where originality is expected". Formsof academic plagiarism vary in their degree of obfuscation ranging from unaltered copies(copy paste), to slightly altered forms of plagiarism, such as interweaving text passages from multiple sources (shake paste), to disguised forms of plagiarism, including paraphrases, translations, and idea plagiarism, and even the plagiarism ofacademic data. The easily identifiable copy paste- type plagiarism is more prevalent among students, while heavily modified plagiarism is more characteristic of researchers, who have strong incentives to avoid detection by skilfully disguising unoriginal content. Research on plagiarism detection (PD) has yielded mature systems employing text retrieval to find similar documents. These systems reliably retrieve documents containing copied text, but often fail to identify disguised forms of academic plagiarism. As we briefly explain in Section 2, several approaches have been introduced to complement text-matching methods and to improve the detection capabilities for disguised forms of plagiarism. Compared to the many sophisticated text-based retrieval approaches that have been proposed for PD, analyzing images to detect academic plagiarism has attracted little research. In this paper, we examine the use of image similarity detection techniques as a promising method for plagiarism detection when textual similarity is lacking. For our use case, we define 'images' as the visual representations of data, e.g., in the form of bar charts, scatter plots, graphs, etc., as well as of concepts in the form of figures showing the schematic representations of entities and their relations, e.g., flow charts, organigrams, and component diagrams. Our definition also includes photographs and photo-realistic renderings. Images enable conveying much information in a compressed format, and they represent this information differently from the informationconveyed in text. These characteristics makeimages a promising feature to examine whenassessing the semantic similarity present in academic documents. Identifying semantic similarity is crucial for detecting translated plagiarism and idea plagiarism. In some cases, even the plagiarism of data becomes detectable if the data values can be reconstructed from graphs. The paper is structured as follows. In Section 2, we briefly present general PD approaches and previous work on image-based PD. We then begin Section 3 by informing our image-based PD approach through an investigation of image similarities found in documents that have been accused of constituting academic plagiarism. The remainder of Section 3 introduces the methods we developed and subsequently integrated into an adaptive and scalable image-based PDapproach capable of targeting the identified types of image similarity.

2. RELATED WORK

Plagiarism Detection Approaches

Plagiarism detection is a specialized Information Retrieval (IR) task with the objective of comparing an input document to a large collection and retrieving all documents exhibiting similarities above a predefined threshold. PD systems typically follow a two-stage process consisting of candidate retrieval and detailed comparison. For candidate retrieval, the systems commonly employ efficient text retrieval methods, such as n-gram fingerprinting orvector space models. For the detailed comparison, the systems typically apply exhaustive string matching. However, such approaches are limited to finding near copiesof a text. To detect disguised forms of academic plagiarism, researchers have proposed a variety of mono-lingual text analysis approaches employing semantic andsyntactic features, as well as cross-lingual IRmethods. Researchers also showed that hybrid approaches, i.e., the combined analysis of text and other content features, improve the retrieval effectiveness for PDtasks. Alzahrani et al. combined an analysis of text similarity and structural similarity. Gipp and Meuschke showed that the combined analysis of citation patterns and

text similarity improves the identification of concealed academic plagiarism. Pertile et al. confirmed the positive effect of combining citation and text analysis and devised a hybrid approach using machine learning. Recently, Meuschke et al. demonstrated the benefit of analysing the similarity of mathematical expressions and patterns of semantic concepts for improving the identification of academic plagiarism.

Image Analysis for Plagiarism Detection

Few studies have investigated the analysis ofimage similarity for PD. Hardik and Hoda ova use higher degree F-transform to provide a highly efficient and reliable method to identify exact copies of photographs or cropped parts there. However, the method does not consider image alterations aside from cropping. Iwanowski et al. evaluate the suitability of well-established feature point methods, such as SIFT, SURF, and BRISK, to retrieve exact and visually altered copies ofphotographs. Srivastava et al. address the same task using a combination of SIFT features extracted using SIFT and perceptual hashing. Feature point methods identify and match visually interesting areas of a scene. The methods are insensitive to affine image transformations, such as scaling or rotation, and relatively robust to changes in illumination or the introduction of noise. Perceptual hashing describes a set of methods that map perceived content of images, videos, or audio files to a hash value(hash. Images perceived as similar by humans also result in similar hash values, in contrast to cryptographic hashing, in which a minor change in the input results in a drastically different hash value. Thus, thesimilarity of images can be quantified as the similarity of their hash values. If image components, such as shapes, are re-arranged,both feature point methods and perceptual hashing often fail. Iwanowski et al. mention that the effectiveness of the feature pointapproaches they tested decreases if the test images consist of multiple sub-images. We also observed this limitation in our tests. Forexample, the two compound images shown in Figure 10 in Appendix A consist of six and four sub-images, respectively. The image in the later document omits two of thesub-images present in the compound image from the source document. Applying the combination of SIFT feature extractor and MSAC feature estimator to compare these two compound images correctly identifies a high similarity between the two sub-images at the top in both compound images, but does not establish a similarity for the other sub-image pairs.

Comparing Images for Document Plagiarism Detection

The paper presents results of research oriented towards an application of image processing methods into document comparisons in view of their application intoplagiarism-detection systems. Among allimage processing methods, the feature-point ones, thanks to their invariance to various image transforms, are best suited for computing image similarity. In the paper various combination of feature point detectors and descriptors are investigated as potential tool for finding similar images in document. The methods are tested on the database consisting of scientific papers containing 5 well known image processing test images. Also, an idea is presented in the paper how the algorithms computing the image similarity may extend the functionality of plagiarism detection systems.

Reducing Computational Effort for Plagiarism Detection by using Citation Characteristics to Limit Retrieval Space

This paper proposes a hybrid approach to plagiarism detection in academic documents that integrates detection methods using citations, semantic argument structure, and semantic word similarity with character- based methods to achieve a higher detection performance for disguised plagiarism forms. Currently available software for plagiarism detection exclusively performs text string comparisons. These systems find copies, but fail to identify disguised plagiarism, such as paraphrases, translations, or idea plagiarism.Detection approaches that consider semanticsimilarity on word and sentence level exist and have consistently achieved higher detection accuracy for disguised plagiarism forms compared to character-based approaches. However, the high computational effort of these semantic approaches makes them infeasible for use in real-world plagiarism detection scenarios. The proposed hybrid approach uses citation- based methods as a preliminary heuristic to reduce the retrieval space with a relatively low loss in detection accuracy. This preliminary step can then be followed by a computationally more expensive semantic and character-based analysis. We show that such a hybrid approach allows semantic plagiarism detection to become feasible evenon large collections for the first time.

3. METHODOLOGY

1. User Login
2. Upload Source Files
3. Upload Suspicious Files
4. Upload Source Image

5. Upload Suspicious Image

Modules Description:

User Login: user signup process completed and now clicks on ‘Login’ link user is login

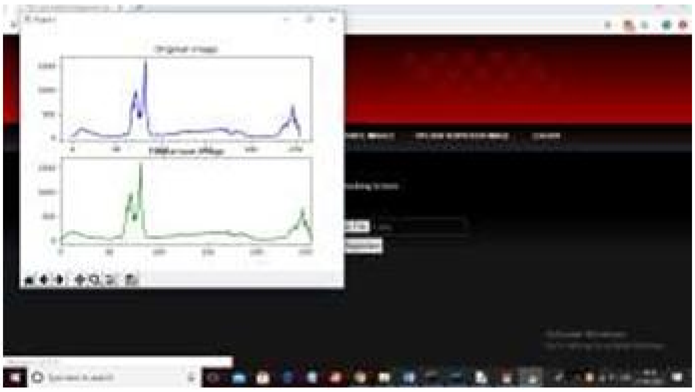
Upload Source Files: Upload Source Files link to load all files from corpus folder and all files are loaded.

Upload Suspicious Files: Upload Suspicious File module to load suspicious file

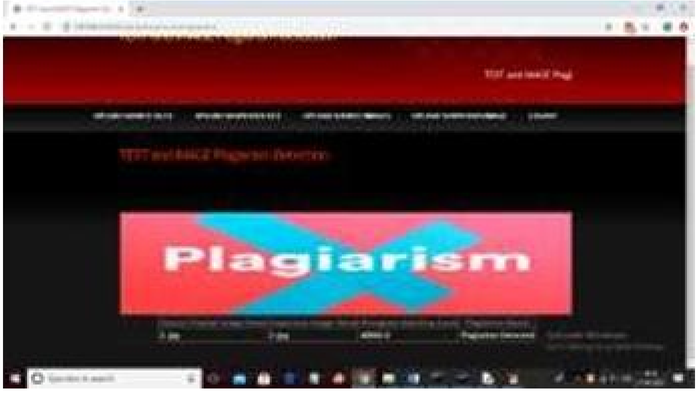
Upload Source Image: Upload Source Images module to upload all images from ‘images’ folder and all database images histogram will be calculated and store in array and whenever we upload new test image then both histograms will get matched.

Upload Suspicious Image: Upload Suspicious Image link to upload some image selecting and uploading ‘112.jpg’ file and then click on ‘Open’ button we can see for database image and uploaded image we generated histogram and we can see there is no match in histogram so no plagiarism will be detected.

4. RESULT AND DISCUSSION



In above screen, it has both original and uploaded image histogram matching 100% so plagiarism is detected



In above screen histogram matching score is 40000 which means all pixels matched so plagiarism is detected in above result.

5.CONCLUSION

We introduced an image-based plagiarism detection approach that adapts itself to forms of image similarity found in academic work. The adaptivity of the approach is achieved by including methods that analyse heterogeneous image features, selectively employing analysis methods depending on their suitability for the input image, using a flexible procedure to determine suspicious image similarities, and enabling easy inclusion of additional analysis methods in the future. To derive requirements for our approach, we examined images contained in the Voila collection. This real-world collection is the result of a crowd-sourced project documenting alleged and confirmed cases of academic plagiarism. From thesecases, we introduced a classification of the image similarity types that we observed. Wesubsequently proposed our adaptive image- based PD approach. Our process integrates perceptual hashing, for which we extended the detection capabilities by including an extraction procedure for sub-images. Since textual labels are common in academic images, we devised and integrated two approaches using OCR to extract text from images and use the textual features for similarity assessments. To address the problem of data reuse, we integrated an analysis method capable of identifying equivalent bar charts. To quantify the suspiciousness of identified similarities, we presented an outlier detection process. The evaluation of our PD process demonstrates reliable performance and extends the detection capabilities of existing image- based detection approaches. We provide our code as open source and encourage otherdevelopers to extend and adapt our approach.

6. REFERENCES

[1] Salha Alzahrani, Vasile Palade, Naomie Salim, and Ajith Abraham. 2011. Using Structural

Information and CitationEvidence to Detect Significant Plagiarism Cases in Scientific Publications. JASIST 63(2) (2011).

[2] Salha M. Alzahrani, Naomie Salim, and Ajith Abraham. 2012. Understanding Plagiarism Linguistic Patterns, TextualFeatures, and Detection Methods. In IEEE Trans. Syst., Man, Cybernet. C, Appl. Rev., Vol. 42.

[3] Yaniv Bernstein and Justin Zobel. 2004. A Scalable System for Identifying Coderivative Documents. In Proc. SPIRE.LNCS, Vol. 3246. Springer.

[4] Teddi Fishman. 2009. –We know it when we see it"? is not good enough: toward a standard definition of plagiarism thattranscends theft, fraud, and copyright. In Proc. Asia Pacific Conf. on Educational Integrity.

[5] Bela Gipp. 2014. Citation-based Plagiarism Detection - Detecting Disguised and Cross-language Plagiarism usingCitation Pattern Analysis. Springer.

[6] Cristian Grozea and Marius Popescu. 2011. The Coplot Similarity Measure for Automatic Detection of Plagiarism. In Proc. PAN WS at CLEF.

[7] Azhar Hamdi, William Puech, Brahim Ait Es Said, and Abdellah Ait Oakman. 2012. Watermarking. Vol. 2. Intech, ChapterPerceptual Image Hashing.

[8] Petr Hardik and Petra Hoda ova. 2015. FTIP: A tool for an image plagiarismdetection. In Proc. Sopra.

[9] Marcin Iwanowski, Arkadiusz Cacok, and Grzegorz Sarwat. 2016. Comparing Images for Document Plagiarism Detection. In Proc. ICCVG.

[10] Yangqin Jia, Evan Shelhamer, Jeff Donahue, Sergey Karasev, Jonathan Long, Ross Kirchick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proc. Multimedia. H.F. Judson. 2004. The Great Betrayal: Fraud in Science. Harcourt.

[11] Jan Kasprzak and Michal Brandeis. 2010. Improving the Reliability of the Plagiarism Detection System. In Proc. PAN WS at CLEF.

[12] Alex Hrushevsky, Ilya Subsieve, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Proc. NIPS.

[13] Donald L. McCabe. 2005. Cheating among College and University Students: A North American Perspective. IJEI 1, 1 (2005).

[14] Norman Meuschke and Bela Gipp. 2013. State-of-the-art in detecting academic plagiarism. IJEI 9, 1 (2013).