A SCALABLE SYSTEM FOR IDENTIFYING SIMILARITY OF DATA FROM DOCUMENTS USING AI

A.Nagalakshmi, Mr.G.Uma Mahesh Mrs.T.Durga Mrs. Aditi Nautiyal

Department of CSE, PRAGATI Engineering College (Autonomous),Surampalem, A.P, India.

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 in a 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 images include 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 the success of this method.**

I 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". Forms of 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 ideaplagiarism, and even the plagiarism of academic 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 skillfully 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 f rms 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 schematicrepresentations 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 information conveyed in text. These characteristics make images a promising feature to examine when assessing 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 beenaccused of constituting academic plagiarism. The remainder of Section 3 introduces the methods we developed and subsequently integrated into an adaptive and scalable image-based PD approach capable of targeting the identified types of image similarity.

II. RELATED WORK

1. 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 or vector space models. For the detailed comparison, the systems typically apply exhaustive string matching. However, such approaches are limited to finding near copies of a text. To detect disguised forms of academic plagiarism, researchers have proposed a variety of mono-lingual text analysis approaches employing semantic and syntactic features, as well as cross-lingual IR methods. Researchers also showed that hybrid approaches, i.e., the combined analysis of text and other content features, improve the retrieval effectiveness for PD tasks. 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 analyzing the similarity of mathematical expressions and patterns of semantic concepts for improving the identification of academic plagiarism.

2. Image Analysis for Plagiarism Detection

Few studies have investigated the analysis of image similarity for PD. Hurtik and Hodakova 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 of photographs. 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 (pHash. Images perceived as similar by Humans also result in similar pHash values, in contrast to cryptographic hashing, in which a minor change in theinput results in a drastically different hash value. Thus, the similarity of images can be quantified as the similarity of their pHash 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 point approaches they tested decreases if the test images consist of multiple sub-images. We also observed this limitation in our tests. For example, 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 the sub-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.

3. 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 into plagiarism-detection systems. Among all image 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 semantic similarity 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 even on large collections for the first time.

4. Optical Character Recognition by Open-source OCR Tool Tesseract: A Case Study.

Optical character recognition (OCR) method has been used in converting printed text into editable text. OCRis very useful and popular method in various applications. Accuracy of OCR can be dependent on text preprocessing and segmentation algorithms. Sometimes it is difficult to retrieve text from the image because of different size, style, orientation, complex background of image etc. We begin this paper with an introduction of Optical Character Recognition (OCR) method, History of Open Source OCR tool Tesseract, architecture of it and experiment result of OCR performed by Tesseract on different kinds images are discussed. We conclude this paper by comparative study of this tool with other commercial OCR tool Transym OCR by considering vehicle number plate as input. From vehicle number plate we tried to extract vehicle number by using Tesseract and Transym and compared these tools based on various parameters. An Evaluation Framework for Plagiarism Detection. We present an evaluation framework for plagiarism detection. The framework provides performance measures that address the specifics of plagiarism detection, and the PAN-PC-10 corpus, which contains 64 558 artificial and 4 000 simulated plagiarism cases, the latter generated via Amazon's Mechanical Turk. We discussthe construction principles behind the measures and the corpus, and we compare the quality of our corpus to existing corpora. Our analysis gives empirical evidence that the construction of tailored training corpora for plagiarism detection can be automated, and hence be done on a large scale. Detecting image plagiarism using hierarchical near duplicate retrieval. Plagiarism in any form is a serious offense especially in academia and industry where integrity and royalty from work is of utmost importance. In this work, a novel hierarchical feature extraction as well as an approximate nearest neighbor search is proposed for detecting plagiarism of images. The proposed scheme is applicable for natural images as opposed to specific image classes reported in a previous work. A comprehensive experimental analysis is provided to illustrate the efficacy of the techniques chosen for the scheme. We demonstrate that the scheme shows a lot of promise for a wide variety of attacks and is amenable to scaling. Comparing and combining Content- and Citation-based approaches for plagiarism detection. The vast amount of scientific publications available online makes it easier for students and researchers to reuse text from other authors and makes it harder for checking the originality of a given text. Reusing text without crediting the original authors is considered plagiarism. A number of studies have reported the prevalence of plagiarism in academia. As a consequence, numerous institutions and researchers are dedicated to devising systems to automate the process of checking for plagiarism. This work focuses on the problem of detecting text reuse in scientific papers. The contributions of this paper are twofold: (a) we survey the existing approaches for plagiarism detection basedon content, based on content and structure, and based on citations and references; and (b) we compare content and citation-based approaches with the goal of evaluating whether they are complementary and if their combination can improve the quality of the detection. We carry out experiments with real data sets of scientific papers and concluded that a combination of the methods can be beneficial. ImageNet Classification with Deep Convolutional Neural Networks We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way SoftMax. To make training faster, we used non- saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called “dropout” that proved to be very effective. We also entered a variant of this model in the ILSVRC- 2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

5. Understanding Plagiarism Linguistic Patterns,Textual Features, and Detection Methods Plagiarism can be of many different natures, ranging from copying texts to adopting ideas, without giving credit to its originator. This paper presents a new taxonomy of plagiarism that highlights differences between literal plagiarism and intelligent plagiarism, from the plagiarist's behavioral point of view. Thetaxonomy supports deep understanding of different linguistic patterns in committing plagiarism, forexample, changing texts into semantically equivalentbut with different words and organization, shorteningtexts with concept generalization and specification, and adopting ideas and important contributions of others. Different textual features that characterize different plagiarism types are discussed. Systematic frameworks and methods of monolingual, extrinsic,intrinsic, and cross-lingual plagiarism detection aresurveyed and correlated with plagiarism types, whichare listed in the taxonomy. We conduct extensive study of state-of-the-art techniques for plagiarism detection, including character n-gram-based (CNG), vector-based (VEC), syntax-based (SYN), semantic- based (SEM), fuzzy-based (FUZZY), structural-based (STRUC), stylometric-based (STYLE), and cross- lingual techniques (CROSS). Our study corroborates those existing systems for plagiarism detection focus on copying text but fail to detect intelligent plagiarism when ideas are presented in different words.

III. SYSTEM ANALYSIS

1. Existing System

Today plagiarism due to the possibility of complex searches on the Web has arrived to research significant damages. Designed tools to deal with plagiarism act on the text and ignore images. On the other, images are inseparable part of information presentation that transfers the large volume of information in an article or scientific research, so that may be one of the plagiarism options.

Disadvantages of Existing System:

1. Less Prediction.
2. More time taking process.2 Proposed System

The images contain a very wide range and especially in the computer literature to be found a lot flowchart images, the purpose of this paper is to examine the plagiarism of a paper in terms of used flowchart images plagiarism using Artificial Neural Networks. The proposed system was tested on 44 images of flowchart images public database that was used for CLEF-IP 2012 competitions. These images have been tested both CVC and INRIA methods. The recognition accuracy average of flowchart test images that have not been tampered in terms of structure, nodes and edges in the proposed method with 81.91 percent is indicating the high success of this method and increase of recognition in compared to both CVC and INRIA method.

Advantages of Proposed System:

1. More Prediction.
2. Less time taking process.

IV.IMPLEMENTATION AND RESULT

In this section, we discuss the implementation of our algorithm in java platform. We first use the parsing technique to parse the text into its constituent data. It returns a set of tokens to be used for pattern matching and compare whether two strings are equal or not .

Fig. 1. Snapshot of the output of parse the text matching in a uploaded file.

In above screen LCS score is 1.0 which means 100% matched with corpus file so plagiarism detected and similarly not only this u may enter any text file and get result. Now click on ‘Upload Source Images’ link to upload all images from ‘images’ folder.

V. COMPARISONS WITH RELATED WORK

To the best of our knowledge, previously comparison with multiple files has not be done using the KNN algorithm, which is implemented in our research. Our logical method intends to find the frequency and accurate result from the multiple files using tokenizer function clustering method. Beside this some more functionalities like each word can give the frequency of repetition we achieve in our project which was not implemented previously. We compare our work with some more papers
based on plagiarism which already exist.

1. In, plagiarism using machine learning language they used some algorithm to detect plagiarism but when we use some code or tool on it then it will not work properly. But in our work, we can take any type of paper and can detect plagiarism.
2. Previously there are methods using artificial intelligence technique for plagiarism but always it does not give better performance. So, we tried to implement k-NN, an artificial intelligence method to produce better result.
3. In grammar rule method also readers face problems like delay in getting result and accuracy in result. Our method detects plagiarism accurately and time consumed is also very less.

VI . FUTURE SCOPE

In our project we have implement the program through which we actually know how various plagiarism tools are working. So, we are planning to add more artificial intelligence method to get more accurate result in future. We will try to know the internal functions of each plagiarism detection tools. We are also planning to store or make a list of all the files related to similar field, so that searching method is become easier.

VII .CONCLUSION

We introduced an image-based plagiarism detection approach that adapts itself to forms of image similarity found in academic work. The adaptively of the approach is achieved by including methods that analyze 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 VroniPlag collection. This real-world collection is the result of a crowd-sourced project documenting alleged and confirmed cases of academic plagiarism. From these cases, we introduced a classification of the image similarity types that we observed. We subsequently 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 other developers to extend and adapt our approach.

VIII. REFERENCES

[1] Salha Alzahrani, Vasile Palade, Naomie Salim, and Ajith Abraham. 2011. Using Structural Information and Citation Evidence 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, Textual Features, and Detection Methods. In IEEE Trans. Syst., Man, Cybern. 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 that transcends 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 using Citation Pattern Analysis. Springer.

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

[7] Azhar Hadmi, William Puech, Brahim Ait Es Said,and Abdellah AitOuahman. 2012. Watermarking. Vol.
2. InTech, Chapter Perceptual Image Hashing.

[8] Petr Hurtik and Petra Hodakova. 2015. FTIP: A tool for an image plagiarism detection. In Proc. SoCPaR.

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

[10] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proc. Multimedia.

[11] H.F. Judson. 2004. The Great Betrayal: Fraud in Science. Harcourt.

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

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Proc. NIPS.

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

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