A Hybrid framework for Identifying similarity of text and image data from documents

[1]A.Radhakrishna , [2]Mrs.L.Yamuna ,[3]Sandhyarani U , [4]Siddila Kavitha
Department of CSE, PRAGATI Engineering College(Autonomous),
Surampalem, A.P, India.

ABSTRACT

Identifying plagiarized content is a crucial task for educational and research institutions, funding agencies, and academic publishers. Plagiarism detection systems available for productive use reliably identify copied text, or near-copies of text, but often fail to detect disguised forms of academic plagiarism, such as paraphrases, translations, and idea plagiarism. To improve the detection capabilities for disguised forms of academic plagiarism, we analyze the images in academic documents as text-independent features. We propose an adaptive, scalable, and extensible image-based plagiarism detection approach suitable for analyzing a wide range of image similarities that we observed in academic documents. The proposed detection approach integrates established image analysis methods, such as perceptual hashing, with newly developed similarity assessments for images, such as ratio hashing and position-aware OCR text matching. We evaluate our approach using 15 image pairs that are representative of the spectrum of image similarity we observed in alleged and confirmed cases of academic plagiarism. We embed the test cases in a collection of 4,500 related images from academic texts. Our detection approach achieved a recall of 0.73 and a precision of 1. These results indicate that our image-based approach can complement other content-based feature analysis approaches to retrieve potential source documents for suspiciously similar content from large collections. We provide our code as open source to facilitate future research on image-based plagiarism detection.

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 idea plagiarism, 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 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 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 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 PD approach capable of targeting the identified types of image similarity.

Objective of the project

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 examine 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.

LITERATURE SERVUY

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.

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 the input 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.

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.

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. OCR is 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 discuss the 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.

imPlag: 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 based on 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.

RELATED WORK

2.1 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 [24]. For candidate retrieval, the systems commonly employ efficient text retrieval methods, such as n-gram fingerprinting or vector space models [15, 26]. 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 [2]. 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 [1]. Gipp and Meuschke showed that the combined analysis of citation patterns and text similarity improves the identification of concealed academic plagiarism [5, 16]. Pertile et al. confirmed the positive effect of combining citation and text analysis and devised a hybrid approach using machine learning [20]. Recently, Meuschke et al. demonstrated the benefit of analyzing the similarity of mathematical expressions [17] and patterns of semantic concepts [18] for improving the identification of academic plagiarism. 2.2 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 thereof [8]. 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 [9]. Srivastava et al. address the same task using a combination of SIFT features extracted using SIFT and perceptual hashing [23]. 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) [7]. Images perceived as similar by humans also result in similar pHash values, in contrast to cryptographic hashing, in which a minor change in the input 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. This problem can be solved by decomposing the compound image into sub-images and applying near duplicate detection methods, such as perceptual hashing, as we show in our evaluation (cf. Case 6 in Table 1, Section 4). Figure 1: Comparison of compound images using SIFT+MSAC. The approach can only establish a similarity for some of the sub-images. Feature point methods and perceptual hashing typically also fail to establish meaningful similarities for images primarily containing text, e.g., tables inserted as images. Typically, the feature points for individual letters are matched to multiple letters occurring in different places in the comparison document, which prevents identifying meaningful clusters of matching features. In summary, prior research on image-based PD proposed methods that reliably retrieve exact and cropped image copies and images that underwent affine transformations. These methods focus on photographs, for which they achieve good results even if photo quality is reduced or modified, e.g., by blurring. For images that underwent other modifications, such as rearranging shapes in the image, redrawing components of the image, or for images that consist primarily of text, the proposed methods often fail. Compound images should be split into meaningful sub-images before applying feature point methods or perceptual hashing to achieve the best retrieval performance. Identifying other types of image similarity than the comparably modest alterations detectable with the approaches we presented in this section requires additional use-case specific analysisapproaches

k-gram Matching

Determining textual similarity by analyzing matching word or character k-grams is a well-established IR approach. Numerous PD approaches employ variable-size or fixed-size k-grams [2, 15, 26]. For regular texts, k-grams with lengths corresponding to 3-5 words, i.e., approx. 15-30 characters, are used most frequently [3, 6, 12]. To choose a k-gram size for analyzing text in figures extracted using OCR, two use-case specific factors should be considered. First, images typically contain smaller text fragments, such as labels or bullet points. Second, we extract the text content of images using OCR, which is likely to introduce noise, i.e., wrongly recognized characters. Such recognition errors can significantly reduce the accuracy, especially for word k-gram approaches. To account for the likelihood that incorrectly recognized characters occur, we chose a comparably fine-grained k-gram resolution of three characters. Given the typically sparse presence of text in figures, we retain all k-grams identified for an image as an unordered set that forms the k-gram descriptor of that image. Typically, kgram-based PD approaches that analyze entire documents employ some form of k-gram selection [2, 15, 26]. We form the k-gram descriptor for all images of the reference collection during preprocessing and store the descriptors in the reference database. Currently, our prototype performs pairwise comparisons of the k-gram descriptor of an input image to all k-gram descriptors of the reference collection. To scale the image-based detection approach to very large collections, an additional filtering step can easily be introduced, e.g., by indexing individual k-grams and requiring a minimum k-gram overlap to perform the full comparison of the k-gram descriptors. To quantify the distance d of two k-gram descriptors K1 and K2, we use the set-based distance function

d = K1K2/▲K1∩K2 , in which represents ▲ the symmetric difference.

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 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 138 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.

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] (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]

[3] Yaniv Bernstein and Justin Zobel. 2004. A Scalable System for Identifying Coderivative Documents. In Proc. SPIRE. LNCS, Vol. 3246. Springer. [4]

[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

[5] Bela Gipp. 2014. Citation-based Plagiarism Detection - Detecting Disguised and Cross-language Plagiarism using Citation Pattern Analysis. Springer. [6

[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 Ait Ouahman. 2012. Watermarking. Vol. 2. InTech, Chapter Perceptual Image Hashing. [8]

[7] 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.