Figure Plagiarism Detection based on Textual Features Representation

Taiseer Abdalla Elfadil Eisa
Faculty of Computing. Universiti Teknologi Malaysia UTM
Johor, Malaysia
taiseralfadil@hotmail.com

Naomie Salim[1], and Salha alzahrani[2]
[1] Faculty of Computing. Universiti Teknologi Malaysia, Johor
[2] Department of Computer Science. Taif University, Taif, Saudi Arabia

***Abstract*— in an academic environment, plagiarism is the process of copying someone else's text, idea or data verbatim or without due recognition of the source, which is a serious academic offence. Many techniques have been proposed in the literature for detecting plagiarism in texts, but only a few techniques exist for detecting figure plagiarism. The main problem associated with existing techniques of plagiarism detection is that they are not applicable to non-textual elements of figures in research publications. This paper focuses on detecting plagiarism in scientific figures. Textual-reference representation based figure plagiarism detection techniques are proposed and evaluated, based on existing limitations. The proposed techniques use enhanced feature extraction such as textual features and similarity computation methods such as similarity based on textual-reference of figures. The enhanced feature extraction method was found to be capable of extracting textual references such as captions and description texts. The similarity detection method was capable of categorising a given figure as either plagiarised or non-plagiarised from a source collection of scientific publications, depending on a certain threshold value. Results showed that the proposed technique achieved precision=0.78 and recall=0.67 result in terms of the evaluation measure.**

Keywords*— *Figure; plagiarism detection; Textual-reference based; similarity detection;

I. INTRODUCTION

Identifying plagiarism has long been considered to be a serious academic offence and area of research, particularly plagiarism in scientific publications. Existing techniques have been developed based on detecting plagiarism in texts. Only a few techniques have focused on detecting similarities between non-textual elements such as info-graphic forms, including results of experiments, frameworks or diagram figures, which are widely used to represent the quantitative information in scientific publications [1-3]. Almost all existing techniques for text-based plagiarism detection represent the document as a bag-of-words model to detect simple forms of plagiarism. Recent techniques have introduced different representations before the checking process, in order to obtain the semantics of the text, to detect more intelligent plagiarism cases, such as graphic representation [4]. Structural representation takes the way the words are distributed throughout the document into consideration when detecting plagiarism [5-7] and uses information extraction based ontology [8]. A different approach was proposed by Kent and Salim [9]. In their work, instead of comparing all the texts of the two documents, they introduced the idea of comparing documents based on certain features. Four basic features were proposed to reduce the time consumed in the detection process. These features were: top keywords, first sentence similarity, longest common query and phrase subsequence. Thus, their approach does not entail comparing the entire texts in documents under investigation. Rather, plagiarism is identified by comparing these four basic features across various document sources to detect plagiarism.

This paper addresses figure plagiarism detection and proposes to use textual reference representation of figures to detect the extent of similarity between the figures. Previous techniques lack understanding of a document's content because they focus on surface representation of text. Techniques for automatic information extraction allow for extracting information from texts so that different cases of the same information can be detected. The idea is if we can extract the same information from two texts, and these texts are semantically similar then we can understand this as sign that it might actually be a case of plagiarism.

Concept-based approaches are widely used before for image handling, and deal with the text associated with the image using different methods ranging from the traditional key words to extracting and selecting the appropriate concepts of the image. Using different methods, they are applied to retrieve similar images, for image understanding, image summarization and concept-based indexing [10-12]. Many studies have used the concepts extracted from figure captions [13, 14]. Information in figure captions is also used for document summarization [15, 16].

This paper describes the concept-based approach to detect similar figures, using information extraction techniques and introducing the idea of semantic similarity for figures, based on entity similarities, and it outlines a first method for its computation. The idea of entity similarities is not limited to figure plagiarism detection, but can be applied to find any entities of the scientific publications (e.g. Similar tables or charts) with semantically similar content. The proposed technique is called textual reference representation of figure-based figure plagiarism detection; the method is designed to compute the similarity of an input figure to all figures in the corpus. As a pre-processing step, the textual reference representation extractor identifies information contained in the

978-1-5090-6255-3/17/$31.00 ©2017 IEEE

input figure and all corpus figures. Figure plagiarism detection was performed based on comparison figures using these textual references representation of figures.

Textual reference-based representation aims to understand the figure by extracting the information about the figure from the running text of documents using modified features introduced by Bhatia and Mitra [16]: figure caption, figure reference sentence and sentence close to the reference sentence. Figure reference sentence and sentence close to the reference sentence are combined as a single feature to give the description of the text feature.

Given the information in an input figure, it is compared to all the figures of the corpus, one after the other, based on the information contained (figure-to- figure comparison). The figure similarity is broken down into a combination of feature similarities. Thus, as a first step, captions are compared against each other (caption-to-caption comparison), then the similarity between compared captions pairs is computed (caption-to-caption similarity). The caption similarity comparison is followed by comparing the similarity between the two descriptions (description-to-description comparison) and then examining the overall similarity of the figures, considering the similarities for both features (caption and description). The next section explains each step of these processes.

II. TEXTUAL REFERENCE OF FIGURES

The textual reference of a figure or the concept of a figure gives rich information about figure's structure and content. Therefore, in this method, the modified features introduced by[16] are defined to extract the textual reference representation of the figure, as shown in Table 1.

TABLE 1. TEXTUAL REFERENCE FEATURES OF FIGURE

Features	Descriptions
Figure caption	Sentence(s) that appears along with the figure and gives a short textual description of it.
Figure reference sentence	Sentence that refers to the figure in the running text, e.g., "*In Figure 1, we show …*",
Sentences close to the reference sentence.	The sentences before and after reference text.

III. SIMILARITY DETECTION BASED ON TEXTUAL REFERENCE OF FIGURE

In this method, textual reference-based representation was introduced as a new technique for figure plagiarism detection. In this method, each figure was represented with the textual reference text of the figure, a similarity detection method was performed based on comparison between figures based on the extracted features, as explained in Algorithm 1.

Algorithm 1 Textual reference - based figure plagiarism detection.

Input: Suspicious and source documents
Output: Similarity score between figures based on the similarities between the textual references of suspicious and source figures.
Processes:
- **A.** Textual reference extraction
- **B.** Text pre- processing
- **C.** Semantic comparisons and report.

Compute Caption -To - Caption similarity score
Compute Description text- To -Description text similarity score
Compute overall figure similarity score
End.

A. Textual reference representation extraction

The proposed method extracts the text of the caption and reference sentences based on the key words of the "figure" while considering different styles for this key word. In this case, extracted text can be the figure caption or figure reference text (each contain the key word 'figure'), but a set of criteria was introduced to classify them. Lastly, extraction of close text in the figure reference text is identified, and the related textual are then constructed out of these extractions.

B. Text Pre- processing

The data pre-processing step comprised of three sub-steps, which were text segmentation, stop word removal and word stemming. Text Segmentation divided the text documents into sentences. The technology of stop word removal for deleting meaningless words was used. A stemming algorithm to remove the affixes (prefixes and suffixes) in a word, in order to generate its root word, was also applied. This step extracted the significant words from the text and ignored the remaining words. This may have adversely affected the similarity between figures.

C. Semantic comparisons and report.

In order to measure the similarity score between figures based on the textual reference representation of figures, the method was designed to measure the similarity score between similar features. To perform this, each figure in the corpus was represented in the form of caption and descriptions. Similarity between figures was computed based on similar feature comparisons, such as comparing the figure caption in the suspicious figure against the figure caption in the source and so as to compare the descriptive text of the suspicious figure against the descriptive text of the one in the source. In order to measure the similarity score between the textual reference text of source and suspicious figures based on similar features, the suspicious and original texts were broken down into their constituent words. Similarity detection between the words was calculated using the Jaccard similarity equation [17] . Jaccard is popular as many Plagiarism Detection methods with high-detection results in terms of precision and recall have used Jaccard [7, 18, 19]. This type of similarity measure can help to discover plagiarism in the form of exact and near to exact copying. To overcome the problem of synonym replacement, previous studies widely used the WordNet Thesaurus dataset [20] to handle the use of synonymy during the detection processes[21, 22]. In this work, in order to examine and match overlapping words in the suspected and the original texts, to determine if any words have been replaced, the synonyms of every word in the specified text will be extracted. These synonyms will be considered as the appearance of the word itself, when used to detect plagiarism to discover the cases of use of synonyms or phrase rewriting.

Caption-to-caption similarity is computed based on (1) where $Sus_{caption}$ *is* the caption of suspicious figure and $So_{caption}$ is the caption of source figure.

$$Caption_{Sim}(Sus_{Caption}, So_{Caption}) = \frac{|Sus_{Caption} \cap So_{Caption}|}{|Sus_{Caption} \cup So_{Caption}|} \quad (1)$$

Description-to-description similarity is computed based on (2), where $Sus_{Description}$ *is* the description text of suspicious figure and $So_{Description}$ is the description text of source figure.

$$Description_{Sim}(Sus_{Description}, So_{Description}) = \left|\frac{Sus_{Description} \cap So_{Description}}{Sus_{Description} \cup So_{Description}}\right| \quad (2)$$

The figure similarity score is computed as an average similarity score of caption and description similarity score; the final decision on a figure as plagiarised or plagiarism free will be based on (3). The figure is classified as a plagiarised figure if the similarity score is greater than the threshold values, as explained in (4).

$$\text{Textual reference}_{\text{Sim}}(Sus, So) = \frac{Caption_{Sim} + Description_{Sim}}{2} \quad (3)$$

$$\text{Plagiarism report} = \begin{cases} \text{Plagiarised} & \text{If Textual reference}_{\text{Sim}} \geq \text{Threshold} \\ \text{Plagiarism Free} & \text{Otherwise} \end{cases} \quad (4)$$

IV. EXPERIMENTAL DESIGN AND DATASET

The experiment considered the number of detected plagiarised figures from the source figures, based on the similarity between the textual reference representation of the source and suspicious figures. The experiment was carried out using textual reference representation based on the figure plagiarism corpus which is constructed for the purpose of this research. The advantage of this corpus is that it is was developed from real plagiarised cases collected from the Notice of Violation of IEEE Publication Principles[1], in which a report has been assigned to each plagiarism case. There are no simulated cases and the behaviour scenario of plagiarised figures is completely authentic. The experiments were performed on 315 figures, consisting of 122 source figures, 170 plagiarised figures, and 23 plagiarism-free figures. The textual reference representation of the figure is attached to each figure inside the corpus. The plagiarism in the figures has different types of modification to different degrees. Some parts of the plagiarised figures are exact copies, or there has been simple modification, such as deleting or adding some words or strong modifications such as changed words, paraphrasing or summarizing.

V. RESULTS AND EVALUATION

The evaluation was performed using evaluation measures proposed by [23]. Before performing the experimental evaluation, there was a need to set the threshold value for the similarity score used to determine whether a figure is plagiarised or not. Since the threshold value was used to control the accuracy of detected plagiarised cases, choosing an optimal value was required to ensure a balance: the higher the threshold value, the higher the likelihood of plagiarism[24] . Experiments were conducted with different values of thresholds and the number of true detected (TP), Not-detected cases (FN) and detected cases but not defined in the annotation files (FP) cases were computed, as shown in Fig. 1. From the figure, it can be seen that the true detected cases decreased with increasing the threshold value. On the other hand, wrong detections also decreased. As can be seen in fig.1, the threshold 0.6 has smaller numbers of both true and wrong detection than the threshold of 0.4. Non-detected cases increased when the threshold values were increased. These values were used to calculate the precision, recall and F-measure.

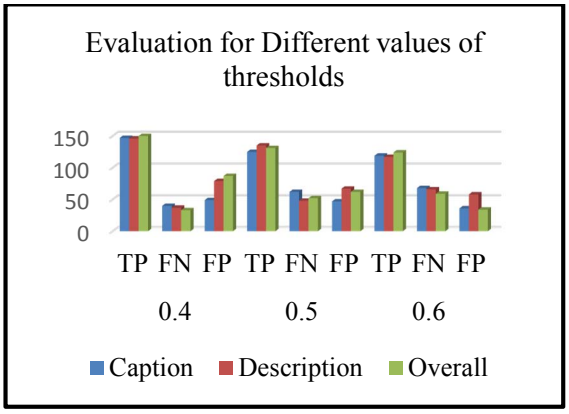

Fig. 1 Evaluation for Different values of thresholds

The standard evaluation measures, precision, recall, and F-measure, were used to evaluate the performance of the textual reference-based figure plagiarism detection techniques. The input to the experiment was the source and the suspicious figures; the experiment aided the comparison of plagiarised figure with the source figures based on the similarities between the related textual reference of the plagiarised and source figures, displaying the source figure and the degree of similarities based on textual matching. In order to investigate which value of threshold can produce a better result, the experiment was run using different threshold values, and precision, recall and F-measure were computed at each threshold value. The better recall results were achieved at a threshold of 0.4, while better precision was achieved at threshold of 0.6. So, the threshold value was set to 0.4 to decrease the wrong detection cases. Beside the method evaluations, an experiment was conducted to investigate the effectiveness of the different textual reference features to detect plagiarised figures. Precision, recall and F-measure were computed for each features. Caption text gives better precision at all threshold values, while description text gives better recall at 0.4 and 0.5 and same result at 0.6. The best results for precision were achieved at 0.6 threshold, while better recall was

[1]http://ieeexplore.ieee.org/search/searchresult.jsp?pageNumber%3D11%26queryText%3DNotice+of+Violation+of+IEEE+Publication+Principles

acquired at 0.4 threshold. The proposed method was able to detect the figures exactly similar to the source figures or modified figures based on related textual reference similarities. It was noticed that, there were some figures that were plagiarised but the proposed method was unable to detect them because the related textual references of these figures were changed by strong modification, such as strong paraphrasing or summarization, even though the proposed method focuses on textual reference modification plagiarism by synonym replacement as well as on copy and paste.

i. CONCLUSIONS

This paper has addressed figure plagiarism detection and proposed the use of semantic similarity between figures based on information extraction techniques. This technique is called textual reference features -based figure plagiarism detection; the method was designed to compute the similarity of a suspicious figure to the source. As a pre-processing step, the textual reference features extractor identifies information related to figures. Comparison between figures is performed based on the related information. The figure similarity is broken down into combinations of feature similarities (caption-to-caption comparison and description-to- description comparison), and the overall similarity of the figure, which considers the similarities for both features (caption and description). The proposed method could detect an exact copy or text with modifications such as adding / removing words, changing word order or replacing words by their synonyms.

ACKNOWLEDGEMENTS

This work is supported by the Malaysian Ministry of Higher Education and the Research Management Centre at the Universiti Teknologi Malaysia under Research University Grant Category Vot: Q.J130000.2528.14H75

REFERENCES

[1] Senosy Arrish, Fadhil Noer Afif, Ahmadu Maidorawa, Naomie Salim, "Shape-Based Plagiarism Detection for Flowchart Figures in Texts". International Journal of Computer Science & InformationTechnology (IJCSIT) 2014. 6(1): p. 113-124. doi: 10.5121/ijcsit.2014.6108.

[2] Rabiu, Idris, and Naomie Salim, "Textual and structural approaches to detecting figure plagiarism in scientific publications". Journal of Theoretical and Applied Information Technology, 2014. 70(2): p. 356-371.

[3] Mohammed Mumtaz Al-Dabbagh, Naomie Salim, Amjad Rehman, Mohammed Hazim Alkawaz, Tanzila Saba, Mznah Al-Rodhaan, and Abdullah Al-Dhelaan, "Intelligent bar chart plagiarism detection in documents". The Scientific World Journal, 2014. 2014: p. 1-11.

[4] Ahmed Hamza Osman, Naomie Salim, Mohammed Salem Binwahlan, Hamza Hentably, Albaraa m. Ali "Conceptual similarity and graph-based method for plagiarism detection". Journal of Theoretical and Applied Information Technology, 2011. 32(2): p. 135-145.

[5] Chow, T.W.S. and M.K.M. Rahman, "Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection". IEEE Transactions on Neural Networks, 2009. 20(9): p. 1385-1402.

[6] Zhang, H. and T.W.S. Chow, "A coarse-to-fine framework to efficiently thwart plagiarism". Pattern Recognition, 2011. 44(2): p. 471-487.

[7] Alzahrani, S.,Salim, N., Abraham, A.,Palade, V. "IPlag: intelligent plagiarism reasoner in scientific publications". in Information and Communication Technologies (WICT), 2011 World Congress on. 2011.

[8] Issa, Hassan, Hose, Katja, Metzger, Steffen, Schenkel, Ralf. "Advances towards semantic plagiarism detection". in Workshop Information Retrieval 2011. LWA: Citeseer.

[9] Kent, C.K. and N. Salim, "Features based text similarity detection". Journal of Computing, 2010. 2(1): p. 53-57.

[10] Roberto Pérez-Rodríguez, Luis Anido-Rifón, Miguel Gómez-Carballa, Marcos Mouriño-García, Architecture of a concept-based information retrieval system for educational resources, Science of Computer Programming, Volume 129, 1 November 2016, Pages 72-91, ISSN 0167-6423, http://doi.org/10.1016/j.scico.2016.05.005.

[11] Ksibi, A., A.B. Ammar, and C.B. Amar. "Enhanced context-based query-to-concept mapping in social image retrieval". in Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on. 2013.

[12] Huynh, N., T.T. Nguyen, and Q. Ho, TeamHCMUS: "A Concept-based information retrieval approach for web medical documents". Proceedings of the ShARe/-CLEF eHealth Evaluation Lab, 2015.

[13] Kahn Jr, C.E. and D.L. Rubin, "Automated semantic indexing of figure captions to improve radiology image retrieval". Journal of the American Medical Informatics Association, 2009. 16(3): p. 380-386.

[14] Guglielmo, E.J. and N.C. Rowe, "Natural-language retrieval of images based on descriptive captions". ACM Transactions on Information Systems (TOIS), 1996. 14(3): p. 237-267.

[15] Futrelle, R.P., "Summarization of diagrams in documents". Advances in Automated Text Summarization, 1999: p. 403-421.

[16] Bhatia, S. and P. Mitra, "Summarizing figures, tables, and algorithms in scientific publications to augment search results". ACM Trans. Inf. Syst., 2012. 30(1): p. 1-24.

[17] Jaccard, P., Etude comparative de la distribution florale dans une portion des Alpes et du Jura. 1901: Impr. Corbaz.

[18] Osman, A. H., Salim, N., Binwahlan, M. S., Twaha, S., Kumar, Y. J., Abuobieda, A.. "Plagiarism detection scheme based on semantic role labeling". in Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on. 2012. Kuala Lumpur: IEEE.

[19] Shuai Wang, Haoliang Qi, Leilei Kong and Cuixia Nu, "Combination of VSM and Jaccard coefficient for external plagiarism detection," 2013 International Conference on Machine Learning and Cybernetics, Tianjin, 2013, pp. 1880-1885. doi: 10.1109/ICMLC.2013.6890902

[20] Miller, G.A., "WordNet: a lexical database for English". Communications of the ACM, 1995. 38(11): p. 39-41.

[21] Al-Shamery, E.S. and H.Q. Gheni, "Plagiarism detection using semantic analysis". Indian Journal of Science and Technology, 2016. 9(1).

[22] Ram, R.V.S., E. Stamatatos, and S.L. Devi, "Identification of plagiarism using syntactic and semantic filters", in Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II, A. Gelbukh, Editor. 2014, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 495-506.

[23] Potthast, M., Stein,B., Barrón-Cedeño, A., Rosso,P. "An evaluation framework for plagiarism detection". in Proceedings of the 23rd international conference on computational linguistics: Posters. 2010. Association for Computational Linguistics.

[24] Jun-Peng Bao,Jun-Yi Shen,Xiao-Dong ,LiuHai-Yan and LiuXiao-Di Zhang, "Semantic sequence kin: a method of document copy detection", in Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004. Proceedings, H. Dai, R. Srikant, and C. Zhang, Editors. 2004, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 529-538.