Plagiarism Detection of Flowchart Images in the Texts

Behnam Hadi
Department of Computer
Malard Branch, Islamic Azad University
Malard, Iran.
behnamhadi@iaumalard.ac.ir

Mohammad Javad Kargar
Department of Computer Engineering,
University of Science and Culture,
Tehran, Iran.
kargar@usc.ac.ir

***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, is 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 are 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.**

Keywords— Plagiarism, high-level semantics, semantic gap, flowchart detection, image mining, image processing

I. INTRODUCTION

Plagiarism can be defined as an unfair appropriation, theft and publication of language, thoughts, ideas or statements of other authors, and redefine they as their main work and in their own name, or stated a thought so that those ideas stay as vague and non-transparent or illegal. This phenomenon, often symmetric in today research and in the creation of information technology and have many influenced in the intensification of consciously or unconsciously plagiarism [1].

Because of its importance, from 2009, Annual international Practical tournament of plagiarism detection evaluation methods, identify the author and related works be held at a conference entitled PAN / CLEF. Since 2012, PAN has developed a web service as TIRA and urged participants, instead of sending the output of its implementation, submitted running software [2].

Plagiarism detection by identifying the source is based on identify the text in massive amounts of documents. In plagiarism Identify without a source, the main focus is on the writers written pattern and structure of the text itself [3].

The third branch of plagiarism identifying has been investigated recently which is operated by the images in texts [4-6] and can be placed in with-source plagiarism identify category.

In [7] plagiarism based on the image is placed in the category of artistic plagiarism, whereas in [8] put it in the category of idea smart plagiarism.

This is a significant gap in the scientific community and requires extensive efforts to deal with this phenomenon. Available software by searching the full text of dissertations and theses restores the similar writings and show the similarity size and the similarity source of information. In this software delete the images. In addition, many of plagiarisms are occur by translating of articles in other languages, especially English and present them with a different title.

Therefore, it can be said that the method of determining the amount of image plagiarism in texts, especially flowchart images is the modern and efficient manner that can be complement text-based plagiarism detection methods and also cover discussions of examining the language of the original article.

II. RELATED WORKS

Martin Potthast the member of the PAN organizing committee and et al [9] for development of an efficient algorithm for plagiarism is defined the plagiarism problem as quadruple of $S = <S_{plg}, d_{plg}, s_{src}, d_{src}>$ that in it s_{plg} is the part of production second document of d_{plg} that occur plagiarism in it and s_{src} is the part of primary document of d_{src} that plagiarism has been done on it.

Plagiarism detector should be able to show an instance of plagiarism to quadruple of $r=<r_{plg}, d_{plg}, r_{src}, d'_{src}$ that in it has been plagiarized the part of r_{plg} from production second document of d_{plg} from the part of r_{src} from the primary document of d'_{src} , r can detect s if and only if equation (1) is satisfied:

$$r_{plg} \cap s_{plg} \neq \emptyset, r_{src} \cap s_{src} \neq \emptyset, \text{and } d'_{src}=d_{src} \quad (1)$$

In PAN-IP-2012, Piroi and et al [10], gave a way to extract the maximum information of image structure in pre-defined text format that it was used in the search for determining the authenticity of the image.

978-1-5386-0420-5/17/$31.00 ©2017 IEEE

plagiarism Identify of the images is tested and attended in a lot of preceding researches. Pan and Wang [11] using Discrete Wavelet Transform(DWT) and two-dimensional Image principal component analysis (2DIMPCA) presented a way to regional plagiarism detection of images. Kour [12] explained frequency based image forgery detection methods, dimensional reduction, the intensity of the image, the correlation between areas of images based on SIFT algorithm, and is expressed a method to identify image counterfeit by focusing on repetitive areas of image based on Fourier transform. Piva [13] done a review of malicious activities on legal medical images and related tools and is emphasis Dempster-Shafer theory for design the combined framework.

Anthony Sutardja and et al [14] manipulate copy-move detection methods in images and have imposed combinations of various descriptors like SIFT, ANMS, Rotation Invariant Box and Highest Corners as Pairwise with each other On 6 different image and compared them with the proposed method. They reached in this conclusions that however the proposed approach in the accuracy and timing is better than some of combination or descriptors alone, but has not the required effectiveness of some of them in diagnosis of resized and multiple copy-move. Bansal and et al [15] are checked and studied methods to plagiarism detection in images of texts and their problems, for instance, images blurred properties that have relatively high computational time and using VSIFT algorithm with Zernike moment by Mohammadian and Poyan [16] that was able to determine the potential geometric transformations.

In [17] using the method of [10], is described a method to flowchart images identify for non-text information retrieval. In the proposed architecture, entries that contain text and images, divided into two layers: text and image. In image layer is used of two point and vector method to extract nodes and flowchart image connections and is used of two node descriptor the names of geometric and BSM descriptor. In [18] Thean and et al Using a technique known as INRIA have the amount of 81.20 percent correct recognition.

Although the information sources to plagiarism identify based on the flowchart images is very low than text-based methods but fortunately a well research studies have been done in flowchart images identifying [18-20]. Lupu and et al [20] is divided criteria for evaluating the effectiveness of the flowchart identification procedures into two general categories: graph-based and query-based methods. Graph-based methods include three MCS, MTM and MXM way and query-based methods include two methods of AR and AP. They explain that their meaning of word query is showing the potential queries which may be a user to a set of flowcharts. In principle, any flowcharts displayed as a series of queries that the answer of each of them is just yes or no. They insist that a patented image recognition especially flowchart Images considerably in terms of processing, different with other images.

Arrish and et al [21] and Maidorawa and et al [22] are described a method for plagiarism detection based on flowchart images. Flowchart image recognition is done by areas identifying. The proposed architecture core is that search plagiarism detection engine, according to the cosine similarity review similarity rate of training image with the other images in database. The system output is number between 0 to 1 that number one means full compliance and zero means non-compliance and numbers between these, mean percentage of compliance of test image to images in the database. The database used consists of 20 flowchart images that have been extracted online resources. They not explain the interpretation of system results.

In [23] some of text-based, citation-based and shape-based plagiarism detection methods have been compared with each other. According to comparison in a copy-paste plagiarism, Text-based plagiarism detection methods have been almost 70 percent whereas citation-based methods inefficient in this regard. About the translated texts plagiarism, text-based methods have been successfully Less than 5 percent, and this value in citation-based method is about 80 percent. In this paper, have not been performed the comparison between these two methods and shape-based method.

Although the articles focused on flowchart images plagiarism identify, to achieve this goal requires further investigation that could be used alone well or to supplement the text-based methods.

III. PROPOSED SYSTEM

The practice of plagiarism detection software is often the first text of all books, articles and dissertations saved in the database and then provide the evaluate possibility and compare the written including: books, articles or dissertations with database content. Based on the algorithms and instructions software work based on it, text possible similarities identified with text databases. Software in start eliminates comparing images in text, and in this manner delete images in the text especially flowchart images that has a lot of information and provide images plagiarism incentives [24-28].

The proposed system has two phases: training and testing as shown fig. 1. They are seen as in train phase used of artificial neural network in learning stage and in the test phase in the recognition stage taken help from the modeling done by this network. Data analysis method and input image similarity detection rate with images in the database is based on the query image correlation rate with each test images and select images with the highest correlation. Correlation levels obtained at this stage report as the tested image plagiarism and the final interpretation is the responsibility of the expert.

The training phase consists of two sub-stages that include:

- features interpretation step
- learning step

In the features interpretation step in first extract examined article images properties and divide in two groups: flowchart images and non-flowchart images as shown in fig. 2. The plagiarism detection operation is performed only on the flowchart images. Then pre-processing is performed on flowchart images and by K-L transform each local features are extracted, after selecting the appropriate number of eigenvectors with the largest eigenvalues, especial images obtained. Finally, each of the special images mapping into

image space and then in the second phase of training, the learn on the mapped especial images is done by artificial neural network that the result of it was to create a model for each of the images.

In test phase first extract the review article images and the operations of features interpretation step is done on the unknown input image and finally detection operation takes place based on belonging to the closest created model in learning step of training phase. In this stage in first images classified in two great categories: flowchart images and non-flowchart images and for non-flowchart images is not performed to determine the amount of plagiarism.

Pre-processing includes thinning operation, delete the text, image histogram equalization, image resize that the recent action use to reduce the computational load of K-L transform.

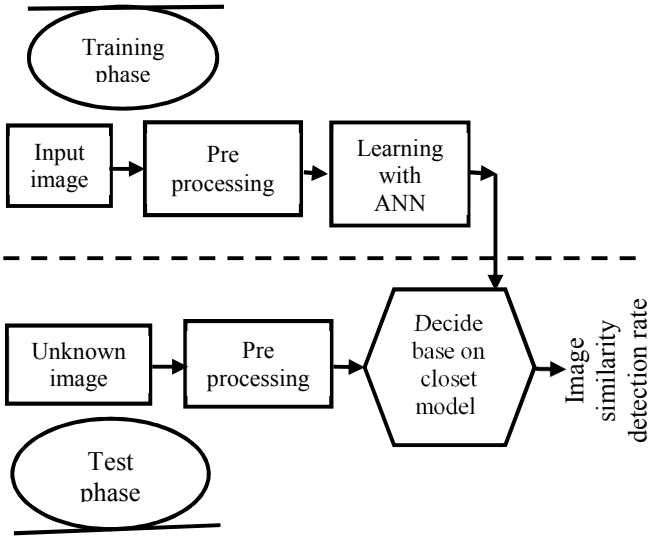

Fig. 1. Our Proposed system

In Fig. 2 shows examples of images in different groups. In flowchart images the text does not matter.

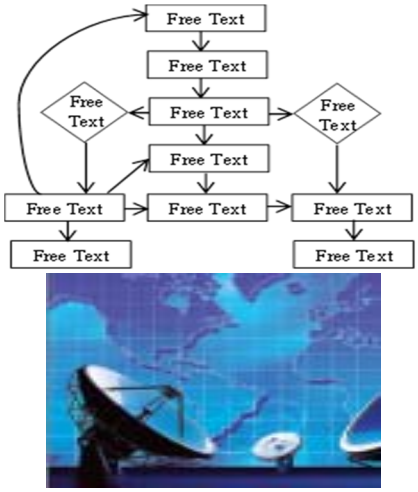

Fig. 2. Examples of two categories of images
Top: flowchart image, bottom: non-flowchart image

IV. EXPERIMENTS

The proposed architecture tested on flowchart images public database that was used for CLEF-IP 2012 competitions. This database contains 150 flowchart cropped photorealistic of approved papers. All images are in binary format and has a flowchart image. Database divide into 50 training images, each with their own text, and 100 testing images. However in CLEF-IP 2012 were evaluated only 44 test images [17].

Fig. 3 shows two examples of the above-mentioned flowcharts images database. In this images the content text is not matter.

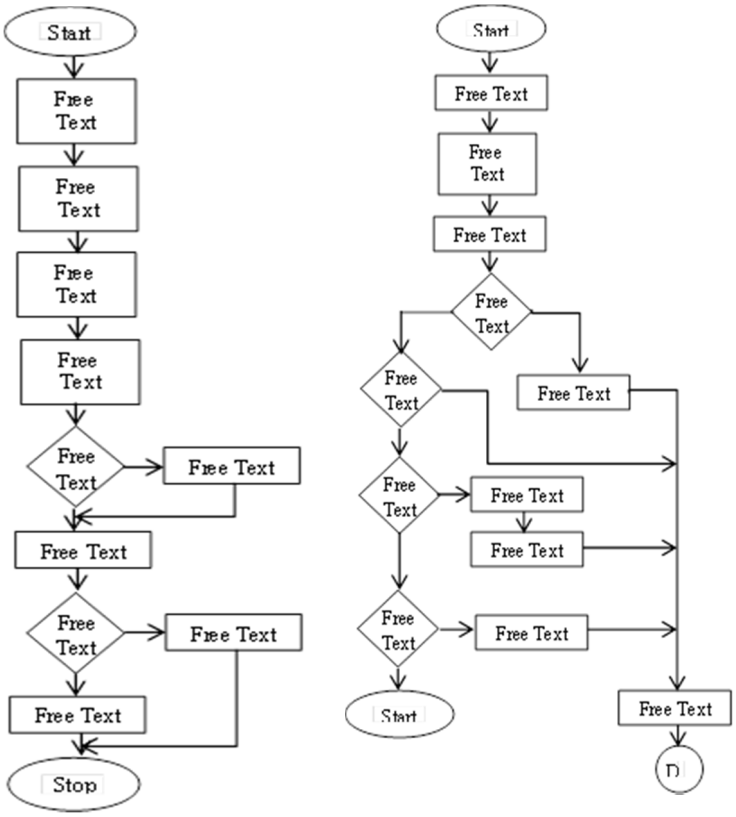

Fig. 3. Two examples of flowchart images of CLEP-IP 2012 competitions database

Table 1 show the results of compares between the proposed architecture with the conclude of several studies that have used the same database.

Table 1: Comparison of proposed architecture with some other methods

Method	Node and Edge Recognition Accuracy	Structure Recognition Accuracy	Recognition Accuracy Average
CVC (Rusiñol et al, [17])	72.50	90.26	81.38
INRIA (Thean et al, [18])	89.09	87.89	88.49
Our Method	91.38	92.24	91.81

As the Table 1 shows the proposed method in terms of flowchart structure recognition and in terms of nodes and edges recognition accuracy of flowchart elements has recognition accuracy than the other two methods. Fig. 4 shows graphically the obtained results.

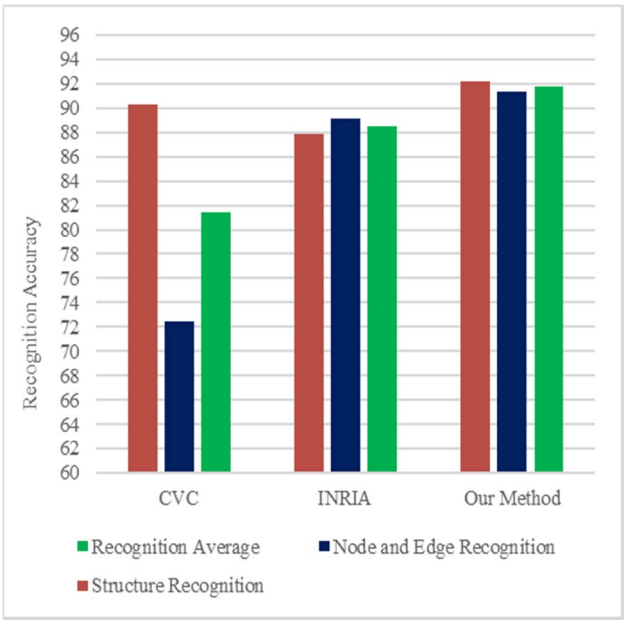

Fig. 4. Compare the proposed approach with some Methods

V. CONCLUSIONS AND RECOMMENDATIONS

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 transfer the large volume of information in an article or scientific research, so that may be one of the plagiarism options. Since 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.

In manipulation flowchart Images diagnosis that include image rotation, image scaling, increase or decrease of flowchart elements, merging elements and dividing elements depending on the amount of manipulation of the image has often different and acceptable precision. Considering that this operation not tested within the systems comprise the proposed system it can work as a researchable work discussed and examined in the future.

REFERENCES

[1] http://www.rahavardnoor.ir/index.php/authors/item/406-serghateadabi, 2014.

[2] Potthast, M., Gollub, T., Stein, B., Rangel F.; Rosso, P., Stamatatos, EU., Stein, B., "Improving the Reproducibility of PAN s Shared Tasks", Information Access Evaluation. Multilinguality, Multimodality, and Interaction: 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, Springer, 2014.

[3] M. Niknam, B. Minaei, "A review on non-textual methods of plagiarism detection," Third International Conference on Applied Research in Computer Engineering and Information Technology. Tehran, Malek Ashtar Industrial University. 2016.

[4] Vipul Bajaj, Sanket Keluskar, Ravi Jaisawal and Prof. Rupali Sawant, "Plagiarism Detection of Images", International Journal of Innovative and Emerging Research in Engineering, Volume 2, Issue 2, 2015.

[5] Ahmadu Maidorawa, Idrissa Djibo, Muhammad Tella, "Plagiarism Detection for Flowchart and Figures in Texts",8th International Conference on Computer Science and Engineering, Malaysia, 2015.

[6] Neha Bansal, Manish Mahajan and Shashi Bhushan, "Comparison of Techniques for Plagiarism Detection in Document Images: A Review", European Journal of Advances in Engineering and Technology, 2(5): 27-31, 2015.

[7] Hermann Maurer, Frank Kappe, Bilal Zaka, "Plagiarism - A Survey", Journal of Universal Computer Science, vol. 12, no. 8, 2006.

[8] Dharmesh Namdev, Jayesh surana, "A Survey Paper on Plagiarism Detection Techniques", International Conference on ICT for Healthcare, 2015.

[9] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, "An evaluation framework for plagiarism detection", in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 997–1005, 2010.

[10] Piroi F, Lupu M, Hanbury A, Sexton A, Magdy W, "Retrieval experiments in the intellectual property domain", In Evaluation Labs and Workshop, Online Working Notes, CLEF 2012.

[11] XZ Pan and HM Wang, "The Detection Method of Image Regional Plagiarism Based DWT and DIMPCA", Advanced Materials Research, 532, 692-696, 2012.

[12] Ramandeep Kaur, "Image Forgery and Detection of Copy Move Forgery in Digital Images: A Survey of Recent Forgery Detection Techniques", International Journal of Computer Applications, Volume 139 – No.5, 2016.

[13] A Piva, "An Overview on Image Forensics", ISRN Signal Processing, 1-22. 2013.

[14] Anthony Sutardja, Omar Ramadan, Yan Zhao, "Forensic Methods for Detecting Image Manipulation Copy-Move", Electrical Engineering and Computer Sciences University of California at Berkeley, 2015.

[15] N Bansal, M. Mahajan and S. Bhushan, "Comparison of Techniques for Plagiarism Detection in Document Images: A Review", European Journal of Advances in Engineering and Technology, 2(5): 27-31, 2015.

[16] Z Mohamadian and A.A. Pouyan, "Detection of Duplication Plagiarism in Digital Images in Uniform and Nonuniform Regions", UK Sim, 2013.

[17] Marçal Rusiñol, Lluís-Pere de las Heras, Oriol Ramos Terrades, "Flowchart recognition for non-textual information retrieval in patent search", Springer, O.R. Inf Retrieval, Volume 17, pp 545–562, 2014.

[18] Thean A, Deltorn J, Lopez P, Romary L, "Textual summarisation of flowcharts in patent drawings for CLEF-IP2012", In: CLEF 2012 Evaluation Labs and Workshop, 2012.

[19] Martin Bresler, Daniel Průša, Václav Hlaváč, "Online recognition of sketched arrow-connected diagrams", International Journal on Document Analysis and Recognition (IJDAR), Springer, 2016.

[20] M Lupu, F Piroi, A Hanbury, "Evaluating Flowchart Recognition for Patent Retrieval", The Fifth International Workshop on Evaluating Information Access (EVIA), Tokyo, Japan, June18, 2013,

[21] Senosy Arrish, Fadhil Noer Afif, Ahmadu Maidorawa and Naomie Salim, "Shape-Based Plagiarism Detection for Flowchart Figures in

Texts", International Journal of Computer Science & Information Technology (IJCSIT) Vol6, No 1, February 2014.

[22] Ahmadu Maidorawa, Idrissa Djibo, Muhammad Tella, "Plagiarism Detection for Flowchart and Figures in Texts", International Journal of Computer and Information Engineering Vol:2, No:8, 2015.

[23] S.A.Hiremath M.S.Otari, "Plagiarism Detection-Different Methods and Their Analysis: Review", International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 1 Issue 7, 2014.

[24] Ahmadu Maidorawa, Idrissa Djibo, Muhammad Tella, "Plagiarism Detection for Flowchart and Figures in Texts", International Journal of Computer and Information Engineering Vol:2, No:8, 2015.

[25] Senosy Arrish, Fadhil Noer Afif, Ahmadu Maidorawa, Naomie Salim, "Shape-Based Plagiarism Detection for Flowchart Figures in Texts," International Journal of Computer Science & Information Technology (IJCSIT) Vol 6, No 1, 2014.

[26] Prajakta Mahendra Ovhal, B.D. Phulpagar, "Plagiarized Image Detection System based on CBIR," International Journal of Emerging Trends & Technology in Computer Science, 2016.

[27] Mohammed Mumtaz, Naomie Salim, and et. All, "Intelligent Bar ChartPlagiarism Detection in Documents" The Scientific World Journal, 2014.

[28] Neha Bansa , Manish Mahajan, "Plagiarism Detection in Document Images using Modified Harris and Belief Propagation", International Journal of Modern Computer Science and Applications (IJMCSA), Vol.3, 2015.