International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Image Document Plagiarism Detection Using DCT Based Local Filtering and Binary Code Hashing

Manpreet Kaur[1], Manit Kapoor[2], Naveen Dhillon[3]

[1, 2, 3]RIET, Punjab, India

Abstract: ***Plagiarism refers to the intellectual property of using the others work in the field of the research. Since easy access to the wide data has become critical problem for researchers, publishers, educational institution. Data is available from the web, large databases and telecommunication network. This paper discusses the technique of the plagiarism for the detection of the copied data in a document image. The approach presented in the paper uses DCT filtering as a source for match study, because each text and image pattern grouped together will yield different frequency pattern. Based on this frequency similarity the detection system finds copy pasted portions of the documents. Performance of system is analyzed on the basis ofnumber of copy paste detected regions and percentage of plagiarism match in scanned content.***

Keywords: Plagiarism, DCT, BOW, Binary pattern, forgery, wavelets

1. Introduction

Plagiarism can be easily seen in every field of the regular life for example in unoriginality of music, artistic creation, specialized articles, pictures, maps and so on. With the advancement of the mobile phones accessibility to the data has been increased which further leads to the innovation of the data. Unoriginality can be characterized of using someone else work without referencing the data with the original author of the data. Since the access to the data has been increased due to access over the electronic material via the web has become the concern for the academic community. It is difficult to define the problem as different institutions and disciplines uses different conventions and has varying tradition. The day may be classified as the common knowledge hence does not requires the referencing to the original author of the data. Plagiarism usually occurs when we use someone else's work, ideas or thoughts as own whether intentionally or unintentionally without giving the proper acknowledgment to the original author. It is important to specify that the plagiarism is not only applicable to the written work such as essay, dissertation, reports or other laboratory results but it is also applicable to plans, music, presentation, projects or any other work for assessment [2].

Scholastic stadium deals with understanding the sensation of the plagiarism identification. Duplicating the content of the others is the offence and needs to be checked for the efficient functioning of the system. Various plagiarism detection frameworks has been developed for the detection of plagiarism in the documents for example turn-it-in [3]. Figure and graphs are disposed before checking for counterfeiting.

Figures and the diagrams are disposed by bringing the look openings so that the individual take the advantage. Copyright infringement framework is used for counterfeit figures and diagrams effectively.

Large numbers of image processing programs such as photoshop are available to create the digital forgeries from one or the multiple images or copying of the various word files into the one file. Detection of plagiarism is used for the detection of manipulation of the image that may act as the evidence for the law enforcement agencies. Law enforcement agencies deal with the general publication document forgery. Plagiarism detection deals with the amount of plagiarism contained in the document and from where the data has been taken

2. Various Categories of the Plagiarism May Include:

- **Accidental:** This type of plagiarism occurs due to the lack of the knowledge regarding the plagiarism and referencing style of the particular institution
- **Unintentional:** As large amount of information is present thoughts and ideas of different persons may be same in spoken or written expressions.
- **Intentional:** Copying the others work without giving proper credit or reference to the original author deliberately.
- **Self-plagiarism:** Publishing the self-work without giving any reference to the original one.
- **Wavelet transformation:** Is one of the most accepted of the time or frequency transformations. Wavelets are indication which are restricted in time and scale and commonly have an unbalanced shape. A wavelet is a waveform of efficiently limited extent that has an average value of zero

The integral wavelet transform is the integral transform defined as:

$$[W\varphi f](a,b) = \frac{1}{\sqrt{|a|}} \int \int_{\infty}^{\infty} \varphi\left(x - \frac{b}{a}\right) f(x)dx$$

The wavelet coefficients c_{jk} are then given by

$$cjk = [W\varphi f(2^{-j}, k2^{-j})$$

Here, $a = 2^{-j}$ is called the binary dilation or dyadic dilation, and $b = k2^{-j}$ is the binary or dyadic position.

3. Literature Survey

Kostoff (Sci Eng Ethics 2006;12(3):543-54) discussed in their work about the finding the similarity between the different documents in the large databases of the research

Licensed Under Creative Commons Attribution CC BY

work. Various fractal articles are analyzed to check the redundancy of the information contained in the article. Various text matching techniques are used abstract the redundancy of the information. After analyzing the full text of the article prediction regarding the redundancy candidate is performed. Small areas of the total articles are analyzed to get the redundant information regarding the article of the candidate. It is the lowest level of the plagiarism that can be seen in the plagiarised text. Violation of the various rules can be seen in the redundant publication of the various articles of the research work.

Various maximization techniques for the reduction of the plagiarism have been discussed. Various papers are analyzed to define the problem in the plagiarism in the documents on the basis of the various parameters. Various metrics that are required to study the performance of the research work are its citations, patents and the publication of the research work for defining the metrics for evaluation of the performance of system by removing the plagiarism from the data

Roig(Psychol Rep. 2005;97(1):43-9)discussed in their work about the analyzing the reused portion of the document in their own work. It is based on the two part study for exploring the plagiarism in the text. Articles and the research work are stored digitally for easy access to them. For finding the plagiarism in the document the new document is compared with the digital database containing the research work by analyzing the strings of the identical data with the help of the software. Modern techniques deals with the confirming the reused data in the document which is identified to be plagiarized . Frequency of the reused content depends from the author to author. Complex methodologies have been designed for the efficient working of the research designs and the procedures followed for the detection of the reused content.

Falagas *et al.***(Arch Immunol Ther Exp (Warsz) 2008;56(4):223-6)**discussed in their work about the "impact factor game" that deals with the observation of the scientific community. Various journals compete to increase their impact factor for gaining the influence in the field of research. Researchers and scientist publish their work in the journals having more impact factor to add value to their research work. Journals competes to come in the top IF journals scientific publication industry works to improve their IF equation by analyzing the disadvantages of the artificial and arbitrary inflation. Quality of the journal depends on the impact factor gained by it. Peer reviews and the editorial of the journals helps to organize the research work in the particular format for ensuring the quality of the research work.

Stark *et al.***(Memory 2007;15(7):776-83)**discussed in their work about the unconscious plagiarism that can be seen in the previous research work. The Various ideas of the others are manipulated to generate the new ides that deals with exposure of initial ideas. New tasks are generated and ideas are improved to get desired work for the research work. Probability of the plagiarism depends on the generative nature of the various the various ideas for the research work. Plagiarism is observed to perform the explicit performance of the new research work done by the author.

Bouville *et al.***(Sci Eng Ethics 2008;14(3):311-22)**discussedin their work about plagiarism. They specify plagiarism as the crime that is performed and need to be analyzed in the academics. The plagiarised content can be labelled separately to distinguish it from the non plagiarised content. This difference is analyzed on the basis of the nature and the importance of the content of the work. Stealing the others content without giving the proper credit is known as the plagiarism and it is the offence in the academic field.

4. Problem Statement

Document image retrieval is the challenging field research work which consists of the continuous interest of the growth and demands of the security for the development of the society. The recognition approach for plagiarism detection is similar to the recognition of the document retrieval system. Survey of the various approaches has being performed and on the basis of which the authors have proposed a new approach performing document image matching using DCT based block correlation and binary code hashing based search scanning. The main objective of is to develop a plagiarism detection technique which provides better results as compared to the technique (Linear Binary Image Matching).

5. Methodology

Select the images document to process for cut and paste process. Selection of the database image set for matching the content for copy paste

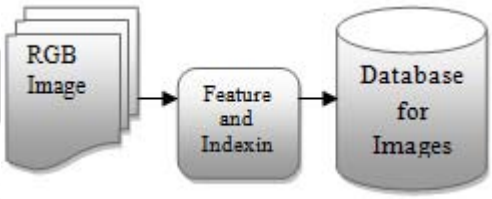

Figure 1: Shows the diagram for image database indexing

Applying similarity and artifact detection using multi-temporal filtering based on DCT analysis and binary hash code.The discrete cosine transform is a method used for converting a signal into straightforward frequency mechanism. It is generally used in image density. Here we build up some simple functions to calculate the DCT and to compress images.The discrete cosine transform of a record of m real numbers t (x'), x' = 0... m-1, is the list of length m given by:

$$R(v)=\frac{\sqrt{2}}{mD(v)} D(v) \in_{y=0}^{m-1} r(y)\cos(zy+1)v\pi/2m \quad v = 0, \ldots\ldots, m$$

Where $D(v)=2^{-1/2}$ for v = 0 or 1 otherwise

Each element of the transform list R (v) is the inner product of the input list r(y) and a basic vector. The stable factors are selected so that the base vectors are orthogonal and normalize. The eight basis vectors for m = 8.

The DCT can be in print as the product of a vector or the input list and the m x m orthogonal matrix whose rows are

Licensed Under Creative Commons Attribution CC BY

the basis vectors. Below visualization shows the DCT transform in various aspects.

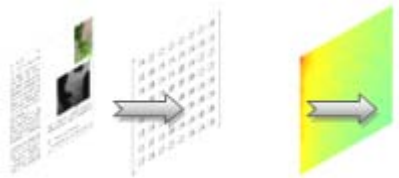

Figure 2: Shows the (left) Original Image; (center) Transform Matrix and (right) Color Map

Perform the feature extraction or segment the image into parts by dividing the document image into 8x8, 16x16, etc. Group data as per user requirement

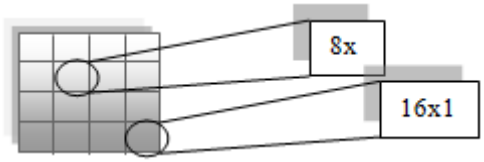

Figure 3: Shows the block based DCT transform for local segment matching

Extract the image block of least probable size. Use the extracted feature for matching with documents database to be under original doc folder.

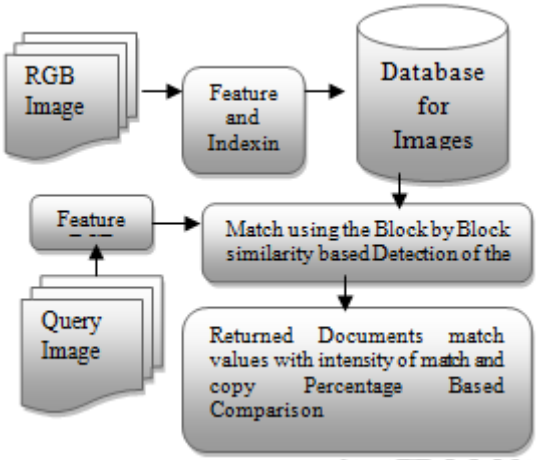

Figure 4: Shows the work flow of the copy paste detection system

After matching, test performance of the processing algorithm on the basis of copy intensity match and plagiarism content percentage. In the proposed system image documents without distortions were used for testing fluency of the algorithm.

6. Results

The following are simulation Results for a number of documents which were forged with text and image copying in order to test and compare the output of the proposed system with that of the previous system in mixed set document evaluation.

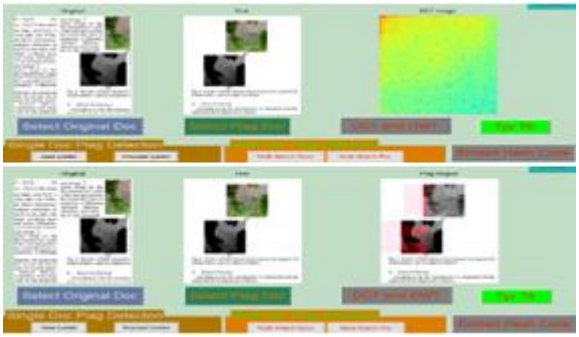

Figure 5: Shows the visual output of the copy paste detection system

The above visual result shows the output of the Plagiarism system which is same for the base system, but the difference is in the match and detection accuracy of the Plagiarism detected documents, the proposed system has a definitive edge on the base system as the number of features extracted from the image documents is high as compared to the base system output, which is shown in below graphical outputs for copy paste detection system.

The test for copy paste detection was done on 9 document images indexed in the database and one document susceptible to be plagiarized. This document was scanned using the proposed model and was matched with dataset documents, the graph below shows the match result of the systems.

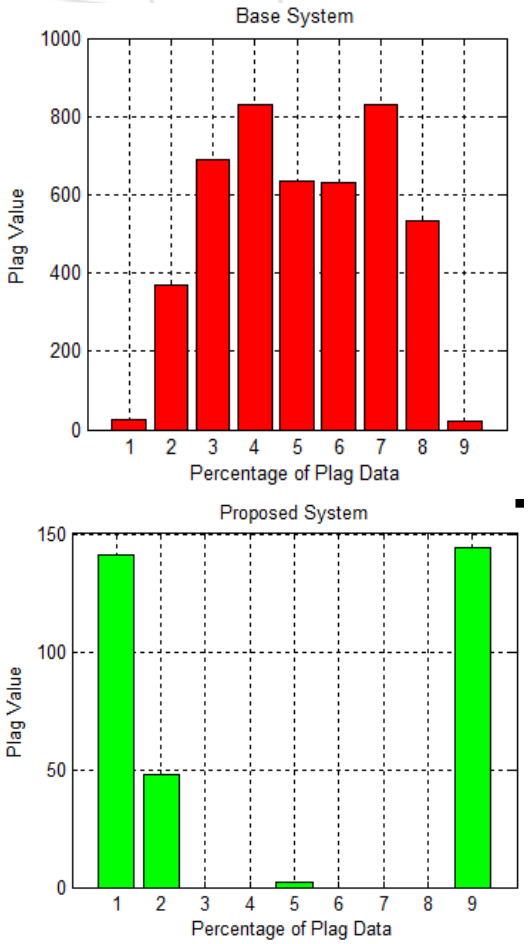

Figure 6: Shows the visual output of the copy paste detection system in graphical form (Red) binary pattern based and (green) DCT based

Licensed Under Creative Commons Attribution CC BY

The above figures show the output of the proposed system for plagiarism match is shown, the maximum match number has reached a value of 96% for the best matched and lowest for 4% the manual evaluation in above the proposed system for documents number 1, 2 and 9. In, the linear pattern match majority of the documents have been shown as the source for copy paste content. The linear match approach shows 50% match in majority documents, when compared with actual content shows false match. In actual case, only the 1 and 2 documents are real sources of content copy from the dataset, for which the DCT approach shows a close match as per the metric evaluation

Table 6.1: Shows Intensity Match Values for Multiple Documents

Doc Name	*DCT Detection*	*Binary Pattern*
1	218594	6932
2	51892	119584
3	100	162350
4	0	344341
5	6500	107563
6	0	45960
7	0	294580
8	400	137740
9	195490	6932

As seen from above tabular representation, the comparison shows contrast between the performance of the binary pattern match and the DCT based match technique.

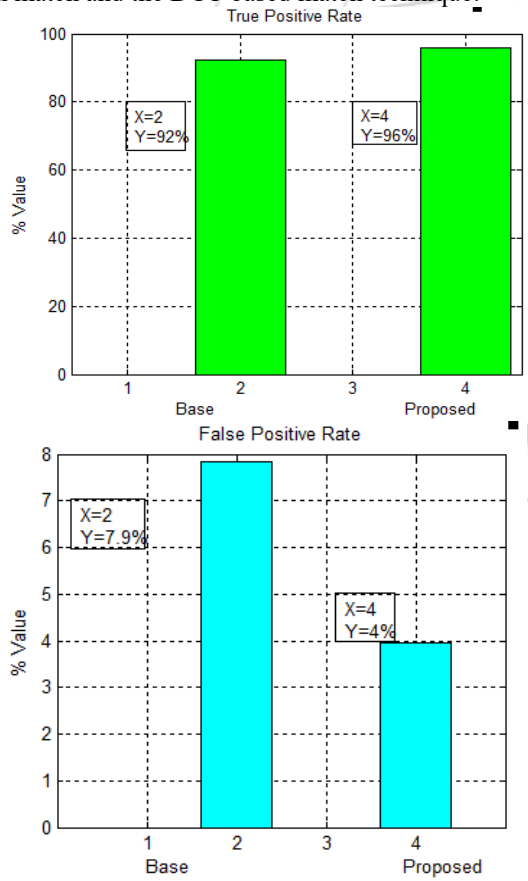

Figure 7: Shows the comparison in graphical form Tpr (Green) and Fpr (cyan) for base and proposed systems

The above figure shows the true positive rate and false positive rate for both the base and proposed system in multiple document copy paste detection, the evaluation shows the increment in Tpr from 92 to 96percent and decrement in Fpr from 8 to 4 percent in proposed w.r.t to base system.

7. Conclusion

The proposed system of work deals with the extraction and match of image documents with the application in detection of duplication work detection for both image and text forgery, the system has different approach from the previous as it extracts the most stable feature points which are not affected by image and text merging done by authors for hiding the copy paste activity, the proof of the system accuracy improvisation is done with comparison of the base technique which only incorporated the text detection and match based plagiarism detection, the overall average increase in efficiency and accuracy of detection in similar documents has proved the robustness of the proposed system over the base.The research presented in this paper proposes a method of what constitutes text and images based plagiarism detection in document images. The system of document copy-paste detection uses the block based approach in the current work study. In future the semantic matching based approach and BOW like feature extraction methods can be used for detail image document copy detection in RGB form and under different environments.

References

[1] H. Maurer, F. Kappe and B. Zaka , “Plagiarism - A Survey,” in Journal of Universal Computer Science (JUCS) , vol.12, no. 8, pp. 1050-1084, 2006.
[2] A.H. Osman, N. Salim and Abuobieda, “Survey of Text Plagiarism Detection,” in Computer Engineering and Applications (CEA), vol. 1, no. 1, pp.37-45, 2012.
[3] U. Ozlem, K. Boris and N. Thade, “Using Syntactic Information to Identify Plagiarism,” in proceedings of Massachusetts Institute of Technology Computer Science and Artificial Intelligence (MITCSAI) , Laboratory Cambridge, USA, pp. 37-44, 2005.
[4] B.G OvGU, “Citation-based Plagiarism Detection – Idea, Implementation and Evaluation,” Germany / UC Berkeley, California, USA, 2010.
[5] A.Juan, C. Nicholas and C. Rafael, “Applying Plagiarism Detection to Engineering Education,” in proceedings of School of Electrical and Information Engineering University of Sydney, NSW, pp. 722-731, 2006.
[6] M. Zimba and S. Xingming, “DWT-PCA (EVD) Based Copy-move Image Plagiarism Detection,” in International Journal of Digital Content Technology and its Applications (IJDCTA), vol. 5, no. 1, 2011.
[7] Z. Ceska, “Plagiarism Detection Based on Singular Value Decomposition,” Springer-Verlag. LNAI 5221, pp. 108–119, 2008.
[8] A. Gandhi and C.V. Jawahar, “Detection of Cut-And-Paste in Document Images,” in proceedings of 12th IEEE International Conference on Document Analysis and Recognition, pp. 653-657, 2013.
[9] A.M. Riad et, al. , “Studying Different Methods for Plagiarism Detection,” in International Journal of Computer Science and Engineering, vol. 2, no. 5, pp. 147-154, 2013.

Licensed Under Creative Commons Attribution CC BY

[10] C.J. Neill and G. Shanmuganthan, "A Web-enabled plagiarism detection tool," in IT Professional, 2004.
[11] G. Whale, "Plague: plagiarism detection using program structure," in Dept. of Computer Science Technical Report 8805, University of NSW, Kensington, Australia, 2008.
[12] J. Y. Kuo, F. C. Huang, C. Hung and L. H. Z. Yang, "The Study of Plagiarism Detection for Object-oriented Programming," in proceedings of Sixth International Conference on Genetic and Evolutionary Computing IEEE, 2012.
[13] P. Lutz, M. Guido and M. Phlippsen, "JPlag: Finding plagiarisms among a set of programs," Fakultä für Informatik Technical Report, Universitä Kalrsruhe, Karlsruhe, Germany, 2000.
[14] M.J. Wise, "Detection of Similarities in Student Programs: YAP'ing may be Preferable to Plague'ing," in ACM SIGSCE Bulletin (in proceedings Of 23rd SIGCSE Technical Symp.), 2002.
[15] M.J. Wise, "YAP3: Improved Detection of Similarities in Computer Programs and Other Texts," SIGCSE, 2006.
[16] M. Potthast et, al., "Overview of the 1st International Competition on Plagiarism Detection," in PAN-09 3rd Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse and 1st International Competition on Plagiarism Detection, 2009, pp. 1-9.
[17] B. Stein et al., "Intrinsic plagiarism analysis," in Language Resources and Evaluation, vol. 45, pp. 63-82, 2011.
[18] S. Eissen and B. Stein, "Intrinsic Plagiarism Detection," in Springer-Verlag ECIR LNCS- 3936, pp. 565–569, 2006.
[19] M. Asim et al., "Overview and Comparison of Plagiarism Detection Tools," Department of Computer Science, Germany & UC Berkeley JCDL, 2011.
[20] B. Stein et al., "Plagiarism analysis, authorship identification, and near-duplicate detection PAN'07," in SIGIR Forum, vol. 41, pp. 68-71, 2007.
[21] S. Gruner and S. Naven, "Tool support for plagiarism detection in text documents," in proceedings of the ACM Symposium on Applied Computing. pp. 776 – 781, 2005.
A. Abraham et al., "Understanding Plagiarism Linguistic Patterns, Textual Features and Detection Methods," Preprint, 2011.
[22] M.J. Wise, "String Similarity via Greedy String Tiling and Running Karp-Rabin Matching," Department of Computer Science, University of Sydney, Australia, 2003.
[23] R. Karp and M. Rabin, "Efficient Randomized Pattern-Matching Algorithms," in IBM Journal of Research and Development, 2007.
[24] S. Tachaphetpiboon, N. Facundes and T. Amornraksa, "Plagiarism Indication by Syntactic-Semantic Analysis," in proceedings of Asia-Pacific Conference on Communications, 2007.
[25] J. Y. Bao, X. D. Shen and H. Y. Liu, "Finding plagiarism based on common semantic sequence model," in proceedings of the 5th International Conference on Advances in Web-Age Information Management, vol.31, no. 29, pp.640–645, 2004.
[26] P. Clough, "Old and new challenges in automatic plagiarism detection," Plagiarism Advisory Service, vol. 10, Department of Computer Science, University of Sheffield, 2003.
[27] S. Gruner and S. Naven, "Tool support for plagiarism detection in text documents," in proceedings of the ACM Symposium on Applied Computing. pp. 776 – 781, 2005.
[28] N. Kang and A.Gelbukh, "PP Checker: Plagiarism Pattern Checker in Document Copy Detection," in LNCS, 4188:661–667. Springer, Heidelberg, 2006.
[29] http://www.doccop.com/
[30] F. Jinan et al., "Designing a Portlet for Plagiarism Detections within a Campus Portal," in Journal of Science (JS), vol. 1, no. 1, pp.83-88, 2005.
[31] B. Mahdian and S. Saic, "Detection of copy–move Plagiarism using a method based on blur moment invariants," in Forensic Science International (FSI), vol.171, no.2, pp. 180-189, 2007.
[32] R. Francisco, G. Antonio and R. Santiago et al., "Detection of Plagiarism in Programming Assignments," in IEEE Transactions on Education, vol. 51, no.2, pp.174-183, 2008.
[33] G. Nathaniel, P. Maria and N. Yiu, "Nowhere to Hide: Finding Plagiarized Documents Based on Sentence Similarity," in proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, NSW, pp. 690-696, 2008.
[34] J. W. Wang, "Fast and robust forensics for image region-duplication Plagiarism," in Automatica Sinica (AS), vol. 35, no. 12, pp.1488-1495, 2009.
[35] Z. Ting and W. Rang-ding, "Copy-move Plagiarism detection based on SVD in digital image," in proceedings of Image and Signal Processing, 2009.
[36] M. Stamm and K. Liu, "Forensic detection of image manipulation using statistical intrinsic fingerprints," in IEEE Transaction on Information Forensics and Security, vol. 5, no. 3, pp. 492–506, 2010.
[37] S. Bravo et al., "Automated detection and localisation of duplicated regions affected by reflection, rotation and scaling," in Image forensics Signal Processing (IFSP), vol. 91, no. 8, pp. 1759-1770, 2011.
[38] M. Sridevi et al., "Copy–move image Plagiarism detection," in Computer Science & Information Technology (CS & IT), vol. 5, no.2, pp. 19-29, 2012.
[39] Z. Ceska, "Plagiarism Detection Based on Singular Value Decomposition," Springer-Verlag. LNAI 5221, pp. 108–119, 2008.
[40] M. Stamm and K. Liu, "Blind forensics of contrast enhancement in digital images," in proceedings of 15th IEEE International Conference on Image Processing, pp. 3112–3115, 2008.
[41] S. Benno, P. Rosso, E. Stamatatos, and Moshe Koppel et al., "Uncovering Plagiarism, Authorship, and Social Software Misuse," in SEPLN Workshop, PAN'09, Donostia-San Sebasti´an, Spain, Universidad Polyt´ecnica de Valencia, 2009.
[42] C. K. Kent and N. Salim, "Web based Cross Language Semantic Plagiarism Detection," in Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing, 2011.
[43] M. Asim, "Using Kohonen Maps and Singular Value Decomposition for Plagiarism Detection," in Third

Licensed Under Creative Commons Attribution CC BY

International Conference on Computational Intelligence, Communication Systems and Networks, IEEE, 2011.
[44] H. Dreher, "Automatic Conceptual Analysis for Plagiarism Detection," in Issues in Informing Science and Information Technology, 2007.

www.ijsr.net
IJSR
ISSN (Online): 2319-7064

Licensed Under Creative Commons Attribution CC BY