Plagiarism Detection Process using Data Mining Techniques

https://doi.org/10.3991/ijes.v5i4.7869

Mahwish Abid(✉), Muhammad Usman, Muhammad Waleed Ashraf
Riphah International University Faisalabad, Pakistan.
mahwish.abid15@gmail.com

Abstract—As the technology is growing very fast and usage of computer systems is increased as compared to the old times, plagiarism is the phenomenon which is increasing day by day. Wrongful appropriation of someone else's work is known as plagiarism. Manually detection of plagiarism is difficult so this process should be automated. There are various tools which can be used for plagiarism detection. Some works on intrinsic plagiarism while other work on extrinsic plagiarism. Data mining the field which can help in detecting the plagiarism as well as can help to improve the efficiency of the process. Different data mining techniques can be used to detect plagiarism. Text mining, clustering, bi-gram, tri-grams, n-grams are the techniques which can help in this process.

Keywords—Plagiarism, Paraphrasing, Data mining, Text mining, MDR, tri-gram, n-gram, Clustering, Similarity, Intrinsic plagiarism, Extrinsic plagiarism

1 Introduction

In this modern time, with the advancement of internet, easy availability of the computers over the globe has made it easy to access other's work which results in plagiarism. Plagiarism is known as the act of using someone else work without the information of author or without giving acknowledge to that corresponding person [1].

With the advancement of technology, use of computers is growing very vastly and it can be seen that they are used everywhere in schools, institutes and industries. More often, assignments of students are submitted in electronic forms. As e-form is easy and suitable for teachers and students as well, but it leads towards the easy opportunity of plagiarism. With the widespread of information over the globe, it is very easy to copy the data from different sources which includes internet, papers, books over the internet, newspapers etc. and paste it in a single work without giving any acknowledge to the sources. These actions lead towards lack of learning in students. So there is a need of detecting the plagiarism to increase and improve the learning of students [2].

Plagiarism can occur in any type of field e.g. novels, program's source codes, research papers and etc. Furthermore, there can occur in numerous situations when

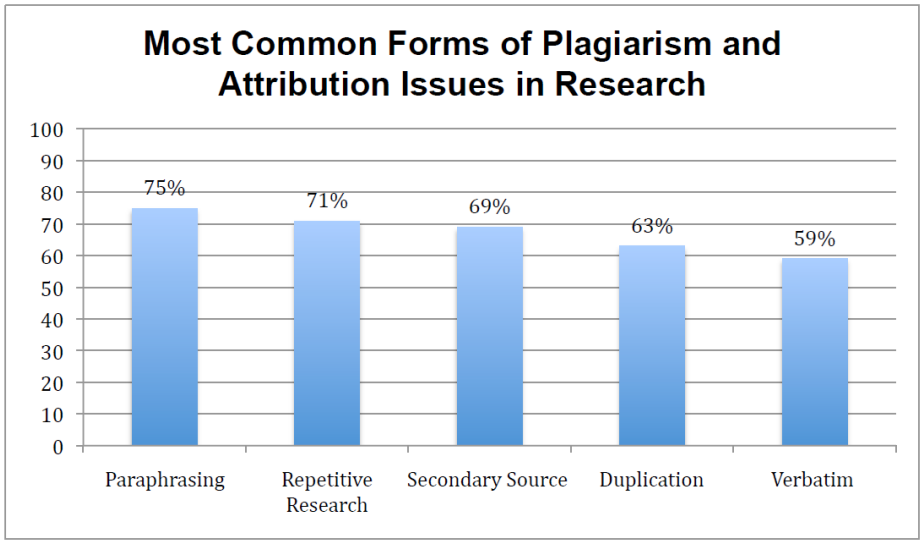

Fig. 1. According to iThenticate survey respondents (Staff, 2013)

students from different institutes copy the data from internet, different books, journals etc. without quoting any reference. Students, sometimes, do this intentionally but mostly they do it unintentionally due to lack of awareness about the usage of resources according to their own need. The issue is not only confined to the written text but programming codes are also include in it. Different small parts of codes are copied from the source and used according to the requirement without referencing the owners [2].

According to a survey which was performed upon plagiarism by university of California from Berkley, they showed that its percentage during the time period of 4 years i.e. 1993-1997 has been increased to 74.4%. And from other studies it was found that above 90.0% students from high schools are included in it [3].

Therefore, plagiarism can be classified into various forms. Some are easily detectable and some are complex. Some of the forms are:

- Coping & pasting: the type in which a single sentence, a whole paragraph or a complete page of written text is copied without any reference [4].
- Re-using existed work: using again the existing work or already written e-data [4].
- Manipulating the text: the type of plagiarism where text is modified and its appearance it changed [5].
- Translating the text: when data is translated from one language to another without giving any reference of the source [5].
- Plagiarizing the idea: one the major form in which someone else's idea is used without acknowledging the owner [6].
- Incorrect citation: citation of unread sources and without giving acknowledge to the other sources from where the data has been read [4].

- Self-plagiarism: the type in which author uses his own previously done work and presenting as new one with any reference of prior work [4].

The plagiarism is difficult to detect manually so it must be automated so that it can be done efficiently. For this purpose, there are different techniques and ways to implement this for example:

- Algorithms to compare documents.
- Crawler to search data from the websites
- Methods using the language-specific structures and much more.

Data mining is one of the field which can be used for this purpose through which relations in existed data can be mined (Hemalatha & Subha, 2014).

2 Literature Review

Plagiarism is defined as the wrongful appropriation or stealing of some other people ideas and make it as own. Stealing or copying of data now a day is becoming very common. Plagiarism detection of copied data originated in 1970's and common methods of Natural language processing (NLP) for detection of copied data introduced in three different techniques namely Grammar-based method, Semantic-based method and Grammar semantic hybrid method [9].

In grammar-based method grammatical structure of the document is maintained and it used a string matching technique for calculating similarity between documents. Semantic based method uses vector space model of the information retrieval technique, and statistical words frequency in document to obtain the vector of the documents, then uses dot product, cousins or other methods to calculate the vectors of two documents. This featured vector is the similarity of the document. This technique isn't effective as it doesn't give the source of the plagiarized data. Grammar semantic hybrid method [10] it improves the detection result of these two methods. It is important and effective to highlight or mark the plagiarized text in the documents in parallel to the similarity results.

In paper [11] author proposed the Longest Common Consecutive Word algorithm, it considers the whole paragraph as a single unit and tracks the words positions. Then by-word comparison is carried out and common words are obtained, this gives the plagiarized version and similarity between documents.

MDR (Match Detect Reveal) is the method in which the document whose plagiarism is going to be checked is first split into the fixed length strings by maintaining a suffix tree. String matching algorithm is used for comparison, and longest common strings can be found in suffix tree. By this, the similarity index and location in the documents can be obtained. This technique is not efficient because it uses the exact words that match and hence making the unclear plagiarized text version [12].

There are different tools which uses web based services and some are standalone applications. Turnitin, article checker and dupli-checker are most common examples of web based services, in these tools except turnitin, other provides the free and online

text bases plagiarism in limited version whereas turnitin supports both intra and extra corpal detection and is not freely available service. Plagiarism Checker X, Copy-Catch, Plagiarism Detector, WORD-Check and CopyFind are standalone application softwares. There are many plagiarism detection approaches which can be used by the applications. Some uses N-gram for improving results in text base. In information retrieval system precision and recall make much senses in calculating accuracy. But as compared to N-gram, bi-gram and tri-gram show much better results than n-grams because, tri--gram shows better precision and bi-gram shows better recall. According to authors they assume that tri-gram sequence matching is effective approach [13].

3 Methodology

a) **Tri-gram and clustering method:** A plagiarism detection process is developed using tri-gram values with the help of comparing the sequences. In this method, the electronic assignments are pre-processed and passed through clustering algorithm. Then tri-gram analysis is performed and similarity results are calculated which are then displayed in the form of percentage [14]**.**
b) **Collecting the data and converting files:** Assignments in electronic form are collected as different three data sets. As all the assignments are different in format so they are converted into a same format.
c) **Pre-processing:** It is an important step to detect plagiarism. In this step data is processed in an appropriate form which can be inputted for detection process. The submitted documents are of different formats including lower and upper case letters. So to remove the sensitiveness, all documents are converted into one format i.e. lower case. Figures, numbers, picture are eliminated.
d) **Constructing the tri-grams:** Three successive word sequences in a line are considered as tri-grams. They are created after processing the assignments. They are formed as shown in Figure 2.
e) **Measuring the similarity:** The comparison is performed on tri-gram structures through the tri-gram comparing method and similarity is calculated. Calculated similarity is depicted in the form of percentage. Greater the percentage shows that similarity is high.
f) **Clustering:** Efficiency of detection process can be increased through clustering technique. For this purpose, K-means algorithm can be applied. The algorithm "K-means" includes a number of advantages for clustering the documents (Sharma, Bajpai, & Mr., 2012).
g) **Stemming:** This technique is used for converting the bag of words to their root words to check that how much this method affects the efficiency of plagiarism. (Jiffriya, Jahan, Ragel, & Deegalla, 2013)

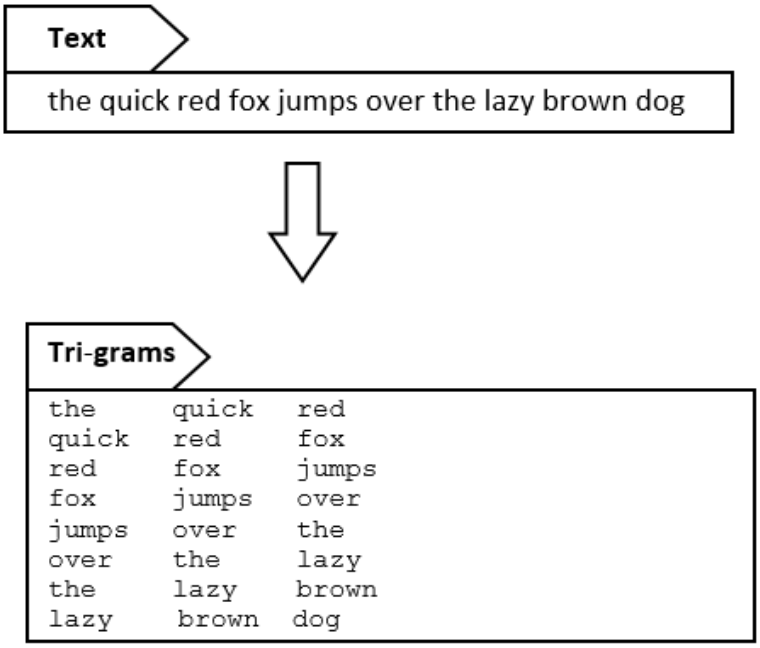

Fig. 2. tri-gram formation [15]

4 Proposed methodology

While using the data mining techniques, plagiarism can be detected easily and efficiently. As various data mining tasks are tradition following, data analyzing technique according to hypothesis, it is a platform to implement adaptable data driven technique that supports the algorithms for detecting the patterns. Basically there are two kinds of data mining techniques which are different as in creating the models or detecting patterns [8].

For this purpose a methodology is proposed in the following:

a) **Collection of assignments:** All the assignments or documents will be collected in electronic format. So that plagiarism can be detected efficiently.
b) **Pre-processing:** Pre-processing is a major step in the process in which all the assignments are converted into a appropriate format. All the assignments collected must be in the same format. Numbers, figure values, pictures and all those things which are not from a-z group should be excluded from the documents.
c) **Classification:** Text classification should be performed to extract and separate the parts of a sentence into alternative words. With the help of this key words from a sentence can be found.
d) **Text analysis:** Further, the data will be passed through the text analyzing step. This process can be repeated, sometimes, according to the need. Moreover differ-

ent text analyzing techniques can be used according to the nature of text and aims of the institutes.

e) **Processing and analyzing the tri-grams:** Sequences of three successive words will be considered as tri-grams in every line. They are created through the cluster of the tri-grams from collection of assignments.
f) **Similarity measures:** Further in the process, comparison is performed upon the sequence of tri-grams created from the processed documents, with the help of sequences comparing methods.
g) **Clustering the plagiarized data:** Clusters are created from the similar tri-grams to calculate the similarity score. Clusters will help in the calculations and will accelerate the process.
h) **Similarity score:** Similarity score will be calculated through the clustering of the similar tri-grams. Similarity will be calculated in the form of percentage. High value of percentage depicts the high similarity score.

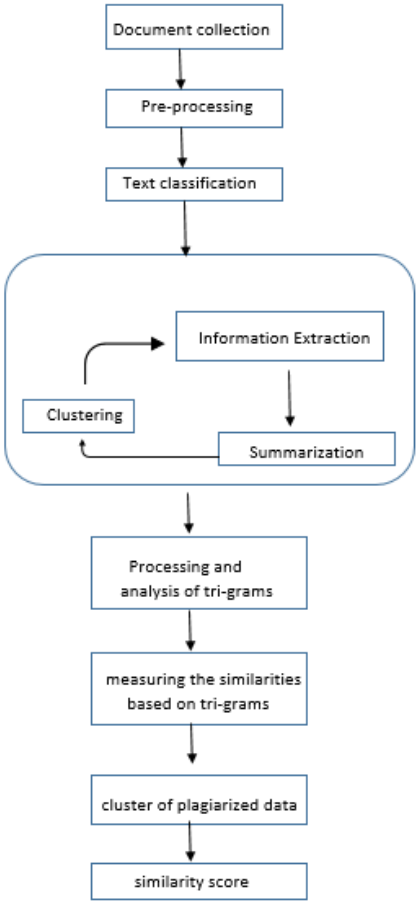

Fig. 3. Proposed Methodology

5 Conclusion

Plagiarism detection process should be automated so that it could be efficient. To enhance the plagiarism detection process, data mining techniques can be used. Here, in this paper, a methodology using data mining techniques is proposed through which, it is thought that the process' efficiency can be improved. Pre-processing and clustering techniques can be used to decrease the overhead of the process. Moreover, similarity score can be calculated through the clusters of plagiarized data so that efficiency can be improved.

6 References

[1] Alzahrani, S. M., Salim, N., Abraham, A., & Senior Member, I. (2012). Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, , 42* (2). https://doi.org/10.1109/TSMCC.2011.2134847

[2] Barron-Cedeno, A., & Rosso, P. (2009). On Automatic Plagiarism Detection Based on n-Grams Comparison. *Springer-Verlag Berlin Heidelberg* , 696-700. https://doi.org/10.1007/978-3-642-00958-7_69

[3] Butakov, S., & Scherbinin, V. (2009). The Toolbox for Local and Global Plagiarism Detection. *Computers & Education , 52* (4). https://doi.org/10.1016/j.compedu.2008.12.001

[4] Clough, P. (2000). Plagiarism in natural and programming languages: an overview of. Department of Computer Science, University of Sheffield.

[5] *Common Forms of Plagiarism*. (2015, may 21). (UNSW sydney) Retrieved september 19, 2017, from https://student.unsw.edu.au/common-forms-plagiarism

[6] El-Matarawy, A., El-Ramly, M., & Bahgat, R. (2013). Plagiarism Detection using Sequential Pattern Mining. *International Journal of Applied Information Systems (IJAIS) , 5*.

[7] Hemalatha, & Subha, M. M. (2014). A STUDY ON PLAGIARISM CHECKING WITH APPROPRIATE ALGORITHM IN DATAMINING. *INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS , 2* (11), 50-58.

[8] Jiffriya, M., Jahan, M. A., Ragel, R. G., & Deegalla, S. (2013). AntiPlag: Plagiarism Detection on Electronic Submissions of Text Based Assignments. *2013 IEEE 8th International Conference on Industrial and Information Systems,* https://doi.org/10.1109/ICIInfS.2013.6732013

[9] Jun-Peng, B., & Shen Jun-Yi, L. X.-D.-B. (2003). A Survey on Natural Language Text Copy Detection. *Journal of Software , 14* (10), 1753-1760.

[10] Roig, M. (2011). Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing.

[11] Sediyono, A., & Mahamud, K. (2008). Algorithm of the Longest Commonly Consecutive Word for Plagiarism Detection in Text Based Document. *Digital Information Management* , 253-259.

[12] Sharma, N., Bajpai, A., & M. R. (2012). Comparison the various clustering algorithms of weka tools. *International Journal of Emerging Technology and Advanced Engineering , 2* (5), 73-80.

[13] Staff, A. (2013, 10 15). *iThenticate study ID's 10 plagiarism formats worthy of attention*. (American Society of Business Publication Editors) Retrieved 09 21, 2017, from https://www.asbpe.org/blog/2013/10/15/ithenticate-study-ids-10-plagiarism-formats-worthy-of-attention/

[14] Tao, W., Xiao-Zhong, F., & Jie, L. (2008). Plagiarism Detection in Chinese Based on Chunk and Paragraph Weight. *In Proceedings of the Seventh International Conference on Machine Learning and Cybernetics.*

[15] *Trigram From Wikipedia, the free encyclopedia*. (2017, September 17). (Wikipedia, the free encyclopedia) Retrieved 09 2017, from https://en.wikipedia.org/wiki/Trigram

[16] Zou, D., Long, W.-j., & Ling, Z. (2010). *A Cluster-Based Plagiarism Detection Method.* Lab Report for PAN at CLEF.

7 Authors

Mahwish Abid, Muhammad Usman, and **Muhammad Waleed Ashraf** are with the Department of Computer Science, Riphah International University Faisalabad, Pakistan.

Article submitted 23 October 2017. Published as resubmitted by the authors 27 November 2017.