IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

SNNO: 2347–3657
Volume 8,Issue2June ,2020

Plagiarism Detection Process using Data Mining Techniques

Mrs. Gangula Pavani , Ms. Noore Ilahi , Ms. Masrath Jahan

Abstract—

Plagiarism is a growing problem in today's digital age because of the widespread availability and increased use of computers and the Internet. The illegal and unjustified taking of another person's plagiarizing someone else's work. Plagiarism detection should be automated because it is a tedious task to perform manually. Plagiarism detection can be accomplished using a number of different methods. There are those that study and work on extrinsic plagiarism, while others focus on intrinsic plagiarism. The efficiency of the process can be increased and plagiarism can be detected with the use of data mining in this area. Plagiarism detection can be achieved through a variety of data mining methods. Helpful methods include text mining, clustering, bi-gram, tri-gram, and n-gram analysis.

1 Introduction

Plagiarism is a widespread problem in today's society because of the widespread availability of computers and the internet. Using another author's words or ideas without properly attributing them is called plagiarism. Without crediting the author or the appropriate correspondent [1]. Computers are becoming ubiquitous as a result of technological progress, serving a wide range of functions in homes, businesses, and other organizations. Students often submit their work electronically. While electronic forms are convenient for both instructors and students, they also provide a heightened risk of plagiarism. With the availability of information all over the world, it is simple to take pieces from other works (online, in print, in books available online, in newspapers, etc.) and combine them into a single piece without properly citing the original authors. Students' failure to learn is a direct result of these behaviors. As a result, identifying instances of plagiarism is essential for enhancing students' education [2]. Plagiarism is a widespread problem in academia and can affect any written work, be it a novel, a computer programmer, a research paper, or anything else. In addition, there are a variety of circumstances in which

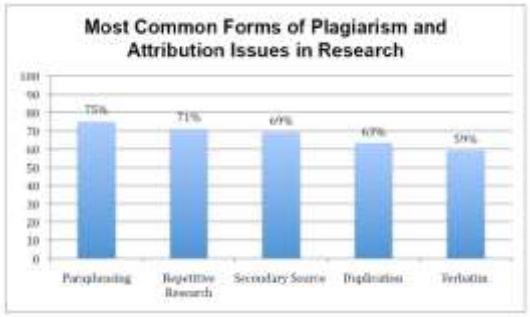


Figure 1. Respondents to an authenticate survey (Staff, 2013) said that students plagiarize when they fail to properly attribute the work of others (including the internet, books, journals, etc.). The students sometimes do this on purpose, although Most of the time, people aren't even aware that they're doing it; they just use more resources than they need. The problem is not limited to words on a page; it extends to computer code as well. It is common practice to steal small pieces of code from the original authors and implement them where needed [2].

1,2,3 Assistant Professor
1,2,3 Department of CSE
1,2,3 Global Institute of Engineering and Technology Moinabad, Ranga Reddy District, Telangana State.

SNNO: 2347–3657
Volume 8,Issue2June ,2020

A survey on plagiarism conducted between 1993 and 1997 at the University of California, Berkeley, found that its prevalence had grown to 74.4% in that time frame. The majority of high school pupils (90.0%) are included, according to prior research [3]. Therefore, there are several kinds of plagiarism. Some of them are simple to spot, while others are more difficult. Examples of available paperwork are:

• Copying and pasting, in which a single sentence, an entire paragraph, or an entire page of text is, lifted wholesale without attribution [4]. By "reusing existing work," we mean recycling or reusing previously created digital content [4]. Plagiarism that involves alterations to the text's appearance is known as "text manipulation" [5]. One example is "translation," which occurs when information is transferred from one language to another without a note of its origin [5]. Plagiarism is the most common method of appropriating the work of others without giving proper credit [6].

Incorrect citations include citing materials that have not been read or failing to properly attribute information borrowed from other sources [4]. The most common form of plagiarism is self-plagiarism, in which the author plagiarizes his or her own work. Providing without citation as if completely original [4]. Because it's so time-consuming and error-prone to do manually, detecting plagiarism is best left to the machines. To achieve this goal, various approaches can be taken, some of which are as follows:

Methods of document comparison algorithms.

• Crawler to search data from the websites

Techniques employing the linguistically-specific tenets, etc.

One subject that can be utilized for this goal is data mining, which allows for the discovery of previously unseen connections within preexisting datasets (Hemalatha & Subha, 2014).

2 Literature Review

Plagiarism is the practice of taking the work of others and passing it off as one's own without giving proper credit. Theft and duplication of data are becoming commonplace. In the 1970s, the first systems to detect instances of copied data emerged. Natural language processing (NLP) strategies for identifying plagiarized content have been introduced in three distinct approaches: a grammatical approach, a semantic approach, and a grammatical-semantic hybrid approach [9].

The grammar-based approach employs a string-matching methodology to determine document similarity while still preserving the document's grammatical structure. After obtaining a document's vector through the information retrieval technique's vector space model and the statistical frequency of words inside the document, a semantically based method may utilize the dot product, cousins, or another way to calculate the vectors of two documents. Here we see the document's similarity represented as a highlighted vector. Since the original source of the plagiarized information is not revealed, this method is ineffective. As a result of combining grammar and semantics, the detection result of these two methods is enhanced [10]. Parallel to the similarity results, it is essential and effective to highlight or otherwise mark the plagiarized text in the documents. The algorithm presented by the author in paper [11] is called Longest Common Consecutive Word, and it treats the entire paragraph as a unit and keeps track of the words' locations inside it. The plagiarized version and the degree of resemblance between papers are then determined by a word-by-word comparison. By keeping a suffix tree, the document whose plagiarism is being detected is first separated into the fixed-length strings using MDR (Match Detect Reveal). Using a string comparison method and the longest common Suffix tree is a great place to look for strings. This allows one to get both the document's similarity index and its location inside it. This method is ineffective because it generates a text that is difficult to distinguish from the original copy due to its reliance on identically worded phrases [12].

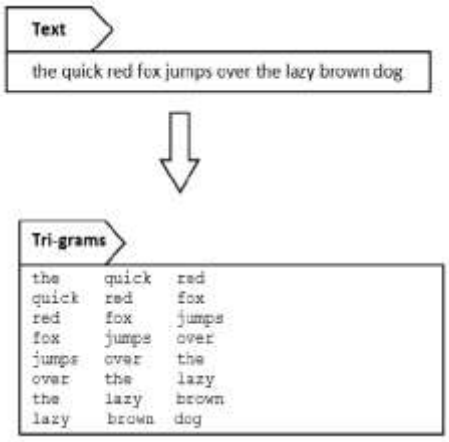
Some of these tools are web-based, while others are more traditional desktop programmers. The most well-known examples of web-based services are Turnitin, Article Checker, and Dupli-Checker; however, only Turnitin charges for its services and offers full support for detecting plagiarism within and outside of its own corpus, whereas the other tools only offer free, online text-based plagiarism detection in a limited version. Separate applications exist for checking for plagiarism, such as Plagiarism Checker X, Copy-Catch, Plagiarism Detector, WORD-Check, and Copy Find. Plagiarism detection software can take a number of different tactics. Some people use N-gram to enhance text base search results. Accuracy calculations make a lot of sense when applied to information retrieval systems, where precision and recall are crucial. However, as compared to N-gram, bi-gram and tri-gram exhibit significantly higher outcomes, particularly in terms of precision (tri-gram) and recall (bi-gram). Tri-gram sequence matching is believed to be an efficient method by the authors [13].

International Journal of
Information Technology & Computer Engineering

SNNO: 2347–3657
Volume 8,Issue2June ,2020

3 Methodologies

The tri-gram and clustering method is used to create a plagiarism detection process by comparing sequences. In this procedure, a clustering algorithm is used to handle the electronic assignments ahead of time. Following this, a tri-gram analysis is run, and the proportion of similarity is reported. [14] Information gathering and file conversion (b) Three distinct data sets are gathered from the electronic assignments. Due to the wide variety of assignment formats, they must all be transformed into one uniform style. When looking for plagiarized content, step c) pre-processing is crucial. Here, the information is transformed into a form usable for the detection procedure. The submitted materials feature a mix of lowercase and uppercase text. Therefore, all papers are changed to lower case to remove any potential for sensitivity. The use of numbers, pictures, and other visual aids has been discontinued. Putting together the tri-grams: a tri-gram consists of three consecutive word sequences in a line. After the assignments have been processed, they are made. Figure 2 depicts their final shape. Tri-gram structures are compared using the tri-gram comparing method, and the level of similarity between them is then determined. The degree of similarity is shown as a percentage when calculated. A higher percentage indicates a higher degree of resemblance. Clustering is a technique that can be used to improve the effectiveness of the detection process. The K-means algorithm is useful for this task. Means is a method for document clustering that has several benefits (Sharma, Bajpai, & Mr., 2012). To determine how much this methodology impacts plagiarism detection, we can employ a process called "stemming," which involves reducing a large vocabulary to its component parts. According to a study (Jiffriya, Jahan, Ragel, & Deegalla, 2013),



4 Proposed methodologies

While applying the data mining approaches, plagiarism can be identified easily and efficiently. Data mining provides a framework for the implementation of flexible data-driven techniques, as many of the activities involved are based on the use of tried and true methods for evaluating data in accordance with preconceived hypotheses. For which the pattern-detection algorithms have support. Data mining methods can be roughly divided into two categories, distinguished by their respective strengths in model construction and pattern discovery [8].

For this goal an approach is proposed in the following:

All assignments and required documentation will be gathered digitally. For the purpose of facilitating reliable plagiarism detection.

b) Pre-processing: This is the stage where all of the assignments are transformed into the correct format. All the assignments collected \smust be in the same format. All non-alphabetic content, such as numeric numbers, graphical representations, and images, should be removed from the papers. Text categorization is necessary for extracting sentence components and categorizing them into different terms. Use this to locate the sentence's most pivotal terms. In addition, the data will undergo a text analysis process. In some cases, this procedure might be repeated. In addition, different text analyzing approaches can be applied depending on the type of material and the goals of the respective institutes.

e) Processing and analyzing the tri-grams: Sequences of three successive words \swill is considered as tri-grams in every line. They emerge when we group trigrams representing a set of tasks together.

Following the processing of texts, a sequence of tri-grams is generated and compared using sequences comparing methods (f) similarity measures.

Clusters are then generated using the tri-gram similarities to determine a similarity score for the plagiarized data. The use of clusters will facilitate the calculations and speed up the procedure.

The similarity score is computed by grouping together tri-grams with similar patterns. Similarity will be calculated in the form of percentage. A high percentage indicates a high degree of resemblance.

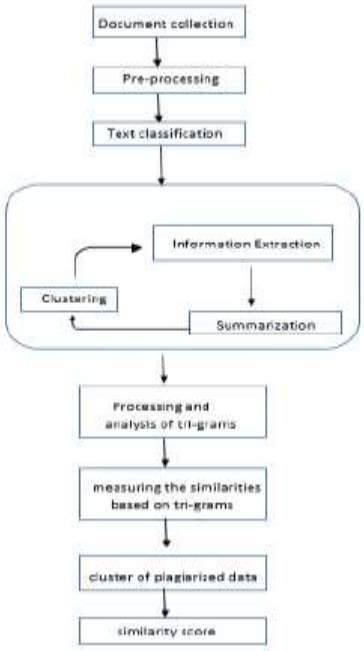


Fig. 3. Proposed Methodology

5 Conclusions

The method of detecting plagiarism needs to be automated so that it can function effectively. Data mining methods can be utilized to improve the originality checking procedure. In this research, we suggest a strategy that makes use of data mining techniques in order to the current level of efficiency is believed to be able to be elevated. Reducing the process's overhead can be achieved through the use of pre-processing and clustering methods. Additionally, efficiency can be enhanced by calculating a similarity score via the plagiarized data clusters.

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] Betake, S., & Scherbinin, V. (2009). The Toolbox for Local and Global Plagiarism Detection. Computers & Education, 52 (4). https://doi.org/10.1016/j.compedu.200 8.12.001

[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-Motorway, A., El-Rally, M., & Bathgate, 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] Jeffrey, M., Johan, 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. 74