IJITCE

ISSN 2347- 3657

International Journal of
Information Technology & Computer Engineering

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

Data Mining Methods for Performing a Plagiarism Check

R ADINARAYANA , K SANTHOSHI

Abstract—

The proliferation and greater usage of computers and the Internet have contributed to the rise of plagiarism in the modern era. Plagiarism is when you take something that doesn't belong to you without permission. Plagiarism detection is a time-consuming process that has to be automated. Several strategies exist for identifying instances of plagiarism. Extrinsic plagiarism is the emphasis of certain researchers, whereas intrinsic plagiarism is the main area of research for others. Data mining in this field may improve efficiency and uncover instances of plagiarism. Several data mining strategies exist for the purpose of identifying instances of plagiarism. Text mining, clustering, bigram, trigram, and n-gram analysis are all useful techniques.

1 Introduction

Due to the increasing availability of computers and the internet, plagiarism has become a major issue in today's society. Plagiarism refers to copying the work of another person without giving due credit. Without giving proper credit to the original author or correspondent [1]. The proliferation of computers in households, companies, and other institutions is a direct effect of technical advancement. Electronic submissions of student work are becoming common. While it's true that electronic forms save time for both teachers and students, they also increase the possibility of plagiarism. These days, it's easy to compile material from several sources (online, in print, in books accessible online, newspapers, etc.) and pass it off as your own without giving credit where credit is due. These actions have a direct impact on students' inability to learn. Therefore, it is crucial for improving students' education to recognize instances of plagiarism [2]. Whether it's a book, a computer program, a research paper, or anything else written, plagiarism is a serious issue in the academic world. Furthermore, there are several situations in which

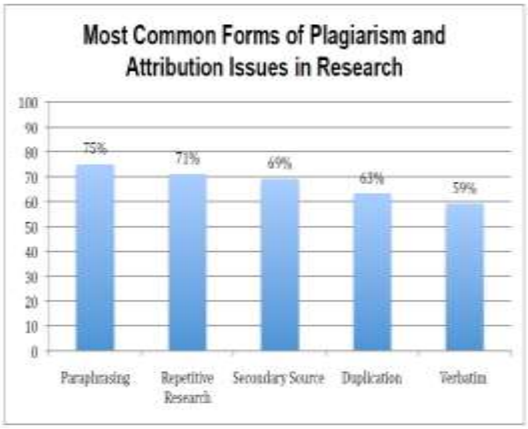


Figure 1. According to a credible poll (Staff, 2013), students plagiarize when they fail to provide adequate credit where credit is due (e.g., when using information from the internet, books, journals, etc.). Sometimes students do this on purpose, but most of the time individuals inadvertently utilize more resources than they need without realizing it. The issue is not confined to text on a page; it also affects code. It's normal practice to lift little snippets of code from other writers and use them when appropriate [2].

Between 1993 and 1997, researchers at the University of California, Berkeley performed a study on plagiarism and discovered that its frequency had increased to 74.4%. According to the literature, [3] the vast majority of high school students (90.0%). As a result, it's clear that there are several forms of plagiarism. The detection of some of them is easy, while that of others is more challenging. Copying and pasting, in which a single line, an entire paragraph, or an entire page of content is copied wholesale without acknowledgement, is one example of the accessible documentation [4]. By "reusing existing work," we refer to digital information that has been recycled or reused [4]. "Text manipulation" is the term used to describe a kind of plagiarism in which the text's physical appearance is changed [5]. One such instance is "translation," which describes the process by which one language's worth of information is converted into another without regard to its original context [5]. The most prevalent kind of stealing someone else's ideas without giving them credit is called "plagiarism" [6].

Not correctly attributing borrowed knowledge or referencing sources that have not been read are examples of incorrect citations. [4]. Self-plagiarism, in which the author plagiarizes his or her own work, is by far the most prevalent kind of plagiarism. Presenting as one's own work without attribution [4]. Plagiarism detection should be automated since doing it manually is tedious and prone to mistakes. Several methods are available for accomplishing this objective.

Approaches to algorithmic document comparison.

- A web information gatherer called a "crawler"

Methods based on the language principles, etc.

Data mining is a field that may help achieve this objective by revealing hidden relationships in current data sets (Hemalatha & Subha, 2014).

2 The Literature Scan

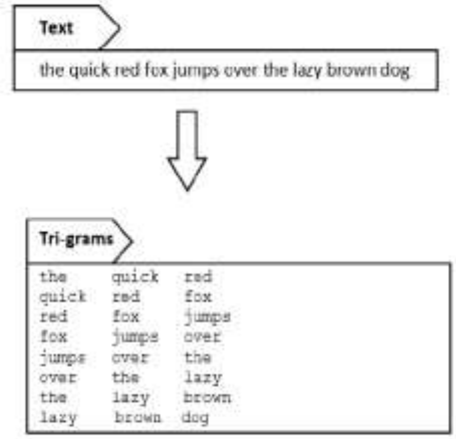
Plagiarism occurs when someone else's ideas or words are presented as one's own without appropriate attribution. Information theft and duplication are more widespread. Systems that could identify instances of copied data appeared in the 1970s. There are three main schools of thought when it comes to using NLP to detect plagiarism: the grammatical method, the semantic approach, and the hybrid grammatical/semantic approach [9].

When comparing documents for similarities, the grammar-based technique uses a string-matching mechanism that doesn't compromise the documents' grammatical structure. A semantically based approach may use the dot product, cousins, or another method to determine the vectors of two documents after first acquiring a document's vector using the information retrieval technique's vector space model and the statistical frequency of words inside the document. In this case, the document's similarity is emphasized as a vector. Ineffective since it does not identify the original author of the pirated material. When grammar and semantics are used together, detection results improve [10]. In tandem with the similarity findings, effectively marking the plagiarized material in the papers is vital. Longest Common Consecutive Word is the author's proposed method in article [11]. It considers the paragraph as a whole and maintains track of the words' positions inside it. Word-by-word comparison then establishes the plagiarized version and the similarity level between works. First, utilizing MDR (Match Detect Reveal), the document being checked for plagiarized content is broken down into the fixed-length strings. One of the best places to search for strings is in the longest common Suffix tree, which may be found by using a string comparison approach. This provides access to the document's similarity index as well as a specific page within it. This strategy fails because it produces content that is hard to tell apart from the original since it uses similarly written terms [12].

There are both web-based and more conventional desktop programmers among these tools. Turnitin, Article Checker, and Dupli-Checker are three of the most popular online services for detecting plagiarism; however, only Turnitin charges for its services and provides full support for detecting plagiarism within and outside of its own corpus. The other tools only provide a limited version of online text-based plagiarism detection for free. Multiple tools, such as Plagiarism Checker X, Copy-Catch, Plagiarism Detector, WORD-Check, and Copy Find, exist specifically for this purpose. There are many approaches that plagiarism detection software might use. N-gram is used by some to improve the outcomes of text-based search engines. Precision and recall are essential in information retrieval systems, therefore accuracy calculations make a lot of sense in this context. Bi-gram and tri-gram both perform better than N-gram, especially in terms of accuracy (tri-gram) and recall (bi-gram). The authors suggest that tri-gram sequence matching is a useful technique [13].

3 Methodologies

By comparing sequences, a plagiarism detection technique may be developed using the tri-gram and clustering approach. In this method, electronic assignments are handled in advance via the use of a clustering algorithm. After that, a tri-gram analysis is performed, and the degree of similarity is indicated. [14] Data mining and format changes (b) There are three different types of information extracted from the electronic homework. Various assignment forms need being converted into a standard format. Plagiarism detection relies heavily on the results of step c) pre-processing. In this step, the data undergoes a transformation that makes it useable in the detection process. There is both lowercase and capitalization in the supplied papers. To avoid any possibility for sensitivity, we have altered all documents to lower case. We no longer make use of any graphical or numerical representations. Putting the tri-grams together: a tri-gram is three word sequences in a row. The assignments are made after the processing is complete. Their ultimate form is seen in Figure 2. The tri-gram comparison technique is used to assess the degree of similarity between two tri-gram structures. When the degree of similarity is computed, it is shown as a percentage. A larger percentage indicates a closer similarity between the two groups. One method that may be utilized to boost detection efficiency is clustering. For this purpose, the K-means algorithm works well. When it comes to document clustering, the Means technique provides a number of advantages (Sharma, Bajpai, & Mr., 2012). We may use "stemming," where a vast vocabulary is broken down into its component components, to assess how much of an influence this strategy has on plagiarism detection. The results of a research (Jiffriya, Jahan, Ragel, & Deegalla, 2013) suggest that



Text

the quick red fox jumps over the lazy brown dog

Tri-grams

the	quick	red
quick	red	fox
red	fox	jumps
fox	jumps	over
jumps	over	the
over	the	lazy
the	lazy	brown
lazy	brown	dog

4 Strategy suggestions

Using data mining techniques, plagiarism may be detected quickly and simply. Data mining lays the groundwork for the use of adaptable data-driven strategies by making use of established procedures for assessing data in light of previously formed assumptions. where the pattern-detection algorithms can function properly. Data mining techniques may be loosely split in two groups, with one focusing on model building and the other on pattern detection [8].

The following is a strategy offered to achieve this end:

We'll be using an electronic system to collect all of your homework and proofreading materials. In an effort to improve the accuracy of plagiarism detection.

b) All assignments are converted into the proper format during this pre-processing phase. All of the submitted work must be presented in the same way. It is recommended that all photos, charts, and other non-textual elements be eliminated from the papers. Extracting sentence parts and placing them in their respective categories requires text classification. You may use this to find the key words in a statement. In addition, textual analysis will be performed on the collected data. This process may need to be repeated in some circumstances. There are a variety of text analysis methods that may be used, each best suited to a certain set of materials and research objectives.

e) The tri-grams are being processed and analyzed. In every line, the sequence of three words swill is regarded a tri-gram. They materialize when we cluster trigrams standing for various jobs.

After the texts have been processed, a sequence of tri-grams is formed and compared using sequences comparing techniques (f) similarity measures.

Next, we use the tri-gram similarities to build clusters and assign a similarity score to the plagiarized information. Calculations will go more quickly and easily thanks to the utilization of clusters.

The similarity score is calculated by clustering tri-grams that have common visual characteristics. The level of resemblance will be expressed as a percentage. When the proportion is large, the degree of similarity is also high.

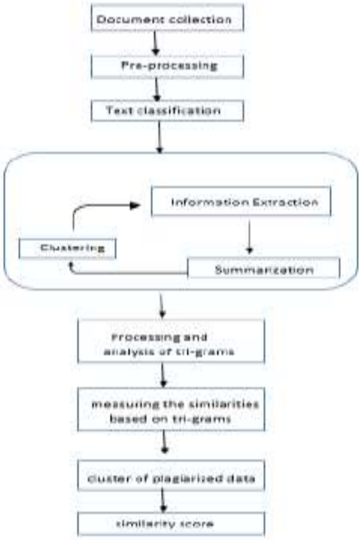


Fig. 3. Proposed Methodology

5 Conclusions

Plagiarism detection processes should be automated for optimal performance. It is possible to enhance the originality testing process by using data mining techniques. We think that the existing level of efficiency can be improved by adopting the method proposed in this study, which employs data mining techniques. Pre-processing and clustering techniques may be used to lessen the burden of the procedure. A similarity score based on the plagiarized data clusters may also be used to improve productivity.

6 References

[1] A. Abraham, I. Senior Member, S. M. Alzahrani, and N. Salim (2012). Learning to Recognize Plagiarism by Knowing Its Language Patterns, Textual Characteristics, and Detection Techniques. Part C: Applications and Reviews, IEEE Transactions on Systems, Man, and Cybernetics, 42 (2). https://doi.org/10.1109/TSMCC.2011.2134847

[2] Reference: Barron-Cedeno, A., & Rosso, P. Regarding n-Gram Comparison-Based Automatic Plagiarism Detection. Page numbers: 697–700 in Springer-Verlag Berlin Heidelberg. The DOI is: 10.1007/978-3-642-00958-7_69.

[3] (2009). Betake, S., & Scherbinin, V. Plagiarism Detection Tools for the World Wide Web. 52(4) of the journal Computers& Education. Obtainable at: https://doi.org/10.1016/j.compedu.2008.12.01

[4] Clough, Patrick. Plagiarism in both human languages and computer languages. University of Sheffield, Computer Science Department.

[5] Different Types of Plagiarism. (2015, may 21). university of new south wales (Sydney) Taken from https://student.unsw.edu.au/common-forms-plagiarism on 2017-09-19.

[6] Anwar El-Motorway, Mohamed El-Rally, and Robert Bathgate (2013). Utilizing Sequential Pattern Mining to Identify Plagiarism. The fifth issue of the International Journal of Applied Information Systems.

7 Hemalatha & M. M. Subha. Examining Plagiarism Detection Algorithms in Data Mining. Research in Computer Applications and Robotics, 2(11), 50-58 (International Journal).

Reference: Jeffrey M., Johan MA, Ragel RG, and Deegalla S. (2013). Plagiarism Detection in Text-Based Electronic Assignments (AntiPlag). ICIInfS.2013.6732013, IEEE 8th International

Conference on Industrial and Information Systems, 2013.

X.-D.-B. Shen, and B. Jun-Peng. 2003. Copy Detection in Text Using Natural Language: a Literature Review. 14(10):1753-1760. Journal of Software.

[10] Roig, M. (2011). How to write ethically and avoid plagiarism, self-plagiarism, and other ethical gray areas.

According to [11] Sediyono, A., and Mahamud, K. (2008). Plagiarism detection algorithm for text-based documents using the longest often occurring sequence of words. Statistics and the Management of Digital Data, 253-259.

Reference: [12] Sharma, N., A. Bajpai, & M. R. Examining the different weka clustering techniques side by side. Two-Fifths of an Issue (25) of an International Journal of New Technologies and Advanced Engineering (7380).