VOLUME-8, ISSUE-1, JANUARY-2019 • PRINT ISSN No 2277 - 8160

Original Research Paper

Computer Science

COMPARATIVE DISCUSSION ON PLAGIARISM CLASSIFICATION AND LEVENSHTEIN DETECTION TECHNIQUES

Nilesh Channawar Gondwana University, Gadchiroli, Maharashtra, India

Dr. S.B.Kishor* Gondwana University, Gadchiroli, Maharashtra, India*Corresponding Author

ABSTRACT Plagiarism is a process of theft of concepts, ideas, words, methodology or results of another person without mentioning the proper acknowledgment, credit or citation. In Plagiarism one person can affirm other work as its own or one can include other hard work as its own without giving the proper citation and credit. Now a day's Plagiarism is a serious concern happening in research field. This paper, presents a systematic review of Plagiarism concept and reviewing on Levenshtein plagiarism detection technique in various text matching field. Levenshtein distance algorithm is a very efficient plagiarism detection technique. We will highlight on the working process of Levenshtein algorithm in detecting Plagiarism.

KEYWORDS : Plagiarism, Levenshtein distance algorithm, Textual Plagiarism, similarity between documents, copyright intrusion.

INTRODUCTION

Plagiarism is now a very serious predicament in professional environments and even in educational systems. Because everyone has access to the Internet, it is easy to use as a source of information. However, copying files from the Internet can be considered plagiarism: anything that can be found on the Internet can come from a book, a research paper or article. These steeling the content may leads to some serious legal issues like copyright intrusion. Now a day's many types of software has discovered .Theft identity software has many types of information where you can use the database like article, book or research papers, Internet or file to file comparison.

The arrival of digital age has increased significantly in the number of digital resources of the web worldwide. Creating such digital resources, their storage and distribution today is simple and straightforward. The rapid increase of this digital company has increased the possibility of copyright infringement and theft. To solve this problem, researchers have seen theft in various languages since 1990, which began with the digital literature duplication detection system [1]. However, the identification program began to detect software misuse through theft in the 1970s [2]. Whereas, many methods and tools are available online for theft identity. However, choosing the best crew identification tool or search detection tool in the best manner is very difficult. This can be due to the lack of proper evaluation environment in the field of plug testing. Plagiarism is the theft of another person's job or idea [3]. It can be done by two ways: (1) To obtain text from specific sources, actions or ideas and (2) present it without the recognition of the text source, work or concept (2). Articles can be stolen in various forms. However, there are often two types of stolen articles, such as: (1) text script plagiarism and (2) source code Plagiarism [4][5]. Theft can occur in a single natural language or in two or more different languages. Many researchers or software vendors are still trying to provide effective methods or tools to detect theft. Generally two types of plagiarism detection techniques are available based on the use of external resources or references [6][5] such as:

1. Intrinsic plagiarism detection: where no external references are used

2. Extrinsic plagiarism detection: where external references are used.

Plagiarism is a Critical problem in the academic world and prevention of it is a very necessary action. Now a days in Information theory and computer science, Levenshtein is a string metric of distance method. This is the method for measuring the distance of the strings. The Levenshtein distance between two strings is given by the minimum number of operations, and that needed to transform one string into the other. Here the operations used are insertion, deletion, or substitution of a single character [7].

In recent years, a series of studies has been done to show that the imitation of tasks, projects and research by other authors has increased rapidly. Due to the research efforts involved in copying other works, the originality of the person has reached its level, which can also achieve the level of academic integrity that plays an important role in the student's life and academic goals. This small academic integrity reflects the desire to achieve the student's goals [8].

PLAGIARISM DETECTION THEORY

The plagiarism can be defined by many definitions but a simple one or an understandable definition would be steeling other work and publishing that work under your name without the knowledge of original author to whom the work belongs. Plagiarism undermines the integrity of education and occurs at all level of scholarship. Now a day's plagiarism has become a problem at academic level. A great way to avoid or minimize the plagiarism is to implement such easy but effective software at academic level like checking the research paper, code etc at academic level before publishing the work[4][8]. Visibility can occur between two identical or different natural languages. Comparison of language can be done by monolingual or textual diversity. Some problems arise in the software for the discovery of stolen articles. The reasons for the similarity between programs can be recognised as follows: One of the incalculable different departments. Sometimes appearance includes metrics, text, detailed features and some grammar tips and some Semantics, program execution, input and output, share information, program dependency and sometimes graphical resemblance.[3][5].

Plagiarism can occur in same or different languages. We can categorise them as monolingual and cross lingual [6].

Monolingual plagiarism detection: In this detection method, same type of languages is taken into account like Germany- Germany or English- English. This type of technique is used vastly now a days. Again it can be intrinsic type and extrinsic type. In intrinsic type external source is not required for detection where the uniqueness is maintained for the writing of an author. In extrinsic type, external sources are required. We have to compare many sources at a time to find out plagiarism [5].

Cross lingual plagiarism detection: It is a different kind of technique for detecting plagiarism. Here different types of languages are used to compare between like English- Germany, Germany- French etc[6][4].

Textual attempts can be classified on the basis of identifying how text features are used to identify documentary masks. Various text features such as Lexical Features, Syntactic Features, Semantic Features, and Structural Features, which can be used to detect similarities between the two documents. These external features are used to identify both cross-lingual plug-in as well as both external and internal [9].

SURVEY ON PLAGIARISM DETECTION ALGORITHM

Plagiarism is one of the most serious morals and problem of education system where a writer is deliberately using any other concept or other key material without accepting its source or submitting it. The problem of theft can be reduced by the integration of literary theft and using other plagiarism detection methods.

Many researchers implement lot of algorithm for detecting plagiarism in academic area. Algorithms that are normally used in plagiarism [10].

The detection software is the ropes, Karp-Rabin Algorithm, Haeckel algorithm, K-gram, string Adaptation algorithm [11]. In [12] the authors describe two algorithms used for testing Efficiency in detection of plagiarism. In [13], the author proposed a system Job properties based on this course Teacher assesses the similarity between the two Submit instead of popular text analysis. The system uses a neural network Create feature-based plagiarism tips Detect and measures the relevance of each function In the assessment [14]. Two popular methods of Levenshtein and Levenshtein Damerau [15] defines the therapeutic distances that can be used Compare the similarities between the two chains Signed each other. Use these distances In various applications of DNA Plagiarism detection analysis.[16] Use Levenshtein's distance to compare words N-gram combined with adjacent similar grams section. Another way [15] Levenshtein Simplified Smith-Waterman algorithm Just like a single algorithm Identify and quantify local similarities In plagiarism detection. [17] Researchers use LCS distance combined with other POS Syntax is used to identify similar local strings Global classification document.[18] Suggested a method based on Eliminate the correct reference for scientific articles To make it a plagiarized paragraph, he can Simulate the actual case of text reuse correctly. [19] Proposed an effective method for the detection of plagiarism Tool, CPLAG, for the programming language code C. The tool evaluates the structure of program C. Based on a set of attributes and execute the binary file The coding of the C code statement Low-cost use in the calculation Recognize the similarity between given C Program. The design of the CPLAG is considered commonly used techniques to avoid detection Counterfeiting ensures effective performance. In addition, heavy calculations are avoided Existing tools for plagiarism detection. [20] It is recommended to use the word2vec template for detection Semantic similarity between words in Arabic .Word2vec is a deep learning method representing a word as a feature of a high value vector Precision. He uses OSAC for training Word2vec template. [21] proposed a plagiarism detection algorithm Approximate string matching Specified in "Copy and Paste", and Increase speed during implementation algorithm. Most of the calculations are the algorithm is omitted by two type's approximate value of power for plagiarism Detect by reducing the accuracy the approximation is acceptable. [22] Proposed a method based on four known methods Model, ie text bag (BOW), Latent Semantic Analysis (LSA), stilometry and Support Vector Machine (SVM).

The most common words (MCW) 25 books Author. The style characteristics of each author are Use in the method by adjusting the LSA Weighting technique. The adjusted LSA method is formed in a new way Verification technology compared Traditional LSA method [23] proposed a fusion method with multiple functions High density based logistic regression model Identification of plagiarized seeds. This method uses a logistic regression model to combine Dictionary function, syntactic function, semantic function Structural features extracted from suspicious text Husband and wife. [24] introduced a different measurement method Semantic similarity between words and them meaning; he proposed a new recognition strategy Plagiarism in user documentation Semantic Web.

PLAGIARISM DETECTION ANALYSIS IN A DOCUMENT

Plagiarism detection method is only one way to secure the plagiarised text in a research paper or in a article or in a book. Some of the basic detection methods are used for plagiarism for a given documents are as follows:

- Tokenized Approach
- Detecting the Same Sentence
- Cementing Based Analysis
- Levenshtein Distance

The first method is to separate the document from the word so that the word can be removed from the document, which reduces the processing time of the document and the complexity of the document. These words, which are reused in the document, are called stop words, in which the words, a, what, where, this, it is. This word type provides text for text. Before any document processing, the stop word should be removed, which reduces the complexity of the document, and the software only transmits the main terms of the competition that defines the document or the study [8]. In some cases, a document may also have delimiters that can be defined as a set of different characters contained in the document to set thresholds. This is another role {} []:; + = And ^ # @! () And sometimes removing such special characters may reduce the complexity of the document. After removing all empty words and special characters in the document. The document contains the main words that have published research concepts. Now that we are working on the document, we need to mark the words that appear in the document, and then use the first method to parse the document, tokenic method. In this approach, words are counted as tokens, which facilitate the execution of other methods.

The second method is to identify the same phrase used to detect close copies in the document. If someone copies the same text from another text and incorporates it, then the time taken to identify the same sentence is longer. This method compares different rows or sentences to identify the same sentence. This method is also used to identify words that have different styles but are used to express similar thoughts. Regardless of which method or technique is used, the number of theft increases when a copy of the text is detected. If the percentage of copied text is greater than threshold, then the process will not be stopped, other parameters will be checked, and we will get the correct results [8].

The third method is based on cement analysis. This method contributes significantly to finding different words with the same meaning. This method is used to search synonyms of tagged words. This method plays an important role in detecting plagiarism; users will use different APIs to get synonyms for the same word. To solve this problem, we use the WORD API, which helps us get the list of synonyms for a particular term. Using such APIs gives access to a large number of words. If no new word exists in the database, the software calls the WORD API and stores the word in the software database. If the same word is again recognized, the software does not need to call the word: The word API always saves a lot of time and money, as if we have crossed the free words limit because we pay for each word Have to do. Case-based analysis also helps to identify words that have been translated and translated many times in different languages so that they can be recognized. It is one of the important processes based on carburizing analysis, which is very useful for people to develop various methods and methods to avoid being caught.

The fourth method is the use of the Levenshtein distance algorithm. The algorithm can be defined as a measure of the similarity between two strings, called source string(s) and destination string (t). Spacing is the number of deletions, inserts, or substitutions required to convert s to t [8].

Levenshtein distance Algorithm:

The Levenshtein distance algorithm is named after the Russian scientist Vladimir Levenshtein. Levenshtein distance algorithm is a very popular and successful plagiarism detection algorithm available in computer science. Now days mostly this algorithm is used to find out the plagiarised text in a required field. The working principle is very simple for the above algorithm[8]. The principle first consider two strings i.e the source string named as 's' and the destination string named as 't'. The algorithm can be simplified as follows:

Step1: If s is "God" and t is "GOD", then LD(s, t) =0, because no transformation are needed. The strings are already identical.

Step 2: s is "GOD" and t is "POD" then LD(s, t) =1, because one substation (change "G" to "P") is sufficient to transform s into t
The greater the Levenshtein Distance the more different the string are. The three main conditions in Levenshtein distance algorithm are as follow –

- The cell immediately above plus 1: d[i-1, j]+1
- The cell immediately to the left plus 1: d[i, j-1]+1
- The cell diagonally above and to the left plus the cost: d [i-1, j-1] +cost.

Levenshitan distance is evaluated the similarity between source strings and target strings. Basic concepts of Levensatin Distance are widely used in computer science, mathematical linguistics, biological sources, molecular biology, DNA analysis and other areas. It can be used to measure music or music similarity. Levenshtein removal is widespread in our daily lives. Levenshtein remove or spell checks removal or edit for program or app fixing. Another possible use of levenshatan distance is in speech recognition and plagiarism identification.

CONCLUSION

Plagiarism in the texts is a matter of concern. Academic community is now the most common area where plagiarism is frequently happening. Mostly plagiarism occurs by common functions like insert, delete or using synonyms for a word. But this type of substitution of work requires lot of string comparison.

In this research paper we provide a broad overview of the plagiarism detection methods and tools used to detect plagiarism. We show various classification and forms of plagiarism that existed in textual data and source code. Here we focus on Levenshtein algorithm for plagiarism detection. Despite the introduction of many methods and tools over the past two decades; we believe that there are still many issues and challenges that need to be addressed. Finally, we highlight a range of research to develop complete and correct plagiarized reviewers for monolingual and multilingual text data as well as source code.

REFERENCES

1. S. Brin, J. Davis, H. Garcia-Molina, Copy detection mechanisms for digital documents, in: ACM SIGMOD Record, Vol. 24, ACM, 1995, pp. 398-409. 24
2. A. Parker, et al., Computer algorithms for plagiarism detection.
3. M. S. Anderson, N. H. Steneck, The problem of plagiarism, in: Urologic Oncology: Seminars and Original Investigations, Vol. 29, Elsevier, 2011, pp. 90-94.
4. N. Charya, K. Doshi, S. Bawkar, R. Shankarmani, Intrinsic plagiarism detection in digital data.
5. Hussain A Chowdhury , Dhruba K Bhattacharyya, Plagiarism: Taxonomy, Tools and Detection Techniques.
6. S. M. Alzahrani, N. Salim, A. Abraham, Understanding plagiarism linguistic patterns, textual features, and detection methods, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42 (2) (2012) 133-149.
7. Zhan Su, Byung-Ryul Ahn, Ki-yol Eom, Min-koo Kang, Jin-Pyung Kim, Moon-Kyun Kim, Plagiarism Detection Using the Levenshtein Distance and Smith-Waterman Algorithm.
8. Nikhil Ghode, Shubham Jadhav, Sampada Moon, Ashmina Khan, Shrutika Bhalkar ,Detecting Plagiarism In Academics Using Levenshtein Distance Algorithm And Semantic Similarity, International Journal on Future Revolution in Computer Science & Communication Engineering ISSN: 2454-4248 Volume: 4 Issue: 3,pp. 471-473
9. C. Barnbaum, Plagiarism: A student's guide to recognizing it and avoiding it.[online].[cit. 2010-12-14] (2009).
10. Yahia jazyah, Open Learning, the Issue of Plagiarism - Efficient Algorithm, International Journal of Computers, ISSN: 2367-8895,Volume 3, 2018
11. P. Clough., "Plagiarism in natural and programming languages: an overview of current tools and technologies", June2000.
12. B.-R. A. Z. Su, K.-Y. Eom, M.-K. Kang, J.-P. Kim,and .-K. Kim,, "Plagiarism detection using the levenshtein distance and smithwaterman algorithm," in ICICIC '08 Proceedings of the 2008 3rd International Conference on Innovative Computing Information and Control. , Washington, DC, USA, 2008, p. 569.
13. V. L. S. Engels, and M. Craig. , "Plagiarism detection using feature-based neural networks.," in Proceedings of the Thirty-Eighth SIGCSE Technical Symposium on Computer Science Education, Covington, Kentucky, March 2007, pp. 34-38.
14. R. M. Federica Mandreoli, Paolo Tiberio, "A document comparison scheme for secure duplicate detection", international Journal of Digital Libraries- Springer-Verlag 2004, Volume 4, Issue 3, November 2004, pp 223–244
15. Z. Su, B. R. Ahn, K.Y. Eom, M. K. Kang, J. P. Kim, and M. K. Kim, "Plagiarism detection using the Levenshtein distance and Smith- Waterman algorithm", 3rd IEEE International Conference on Innovative Computing information And Control, 2008. ICICIC '08. 18-20 June 2008, Dalian, Liaoning, China, DOI: 10.1109/ICICIC.2008.422
16. V. Scherbinin and S. Butakov, "Using Microsoft SQL server platform for plagiarism detection", Stein, Rosso, Stamatatos, Koppel, Agirre (Eds.): PAN'09, pp. 36-37, 2009.
17. C. Barnbaum, "Plagiarism: A Student's Guide to Recognizing It and Avoiding It", 2002. Available at http://www.cpalms.org/Public/PreviewResourc eUrl/Preview/25915
18. Mohtaj S, Asghari H, Zarrabi V. "Compiling a text re-use detection corpus from scientific papers with semi-real cases of plagiarism". International Conference In Asian Language Processing (IALP), 2017 Dec 5 (pp. 227-230). IEEE.
19. ain S, Kaur P, Goyal M, Dhanalekshmi G. "CPLAG: Efficient plagiarism detection using bitwise operations". Tenth International Conference on Contemporary Computing (IC3), 2017 Aug 10 (pp. 1-5). IEEE.
20. Suleiman D, Awajan A, Al-Madi N. "Deep Learning Based Technique for Plagiarism Detection in Arabic Texts". International Conference on New Trends in computing Sciences (ICTCS) 2017 Oct 1 (pp. 216-222).IEEE.
21. Baba K. "Fast plagiarism detection based on simple document similarity". Twelfth International Conference on Digital Information Management (ICDIM), 2017 Sep 12 (pp. 54-58). IEEE.
22. AlSallal M, Iqbal R, Amin S, James A, Palade V. "An Integrated Machine Learning Approach for Extrinsic Plagiarism Detection". 9th International Conference on Developments in eSystems Engineering (DeSE), 2016 Aug 31 (pp. 203-208). IEEE.
23. Kong L, Lu Z, Qi H, Han Z. "High obfuscation plagiarism detection using multi-feature fusion based on Logical Regression model". 4th International Conference on Computer Science and Network Technology (ICCSNT), 2015 Dec 19 (Vol. 1, pp. 355-359). IEEE.
24. Agarwal J, Goudar RH, Kumar P, Sharma N, Parshav V, Sharma R, Srivastava A, Rao S. "Intelligent plagiarism detection mechanism using semantic technology: A different approach". International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2013 Aug 22 (pp. 779-783). IEEE.