A Survey of Plagiarism Detection Strategies and Methodologies in Text Document

Mr. Dnyaneshwar R. Bhalerao
Department of Computer Engineering,
PICT, Pune-411043

Prof. S. S. Sonawane
Department of Computer Engineering,
PICT, Pune-411043

***Abstract*- Research is base for innovation. A number of research articles are available on the internet, either in text or in multimedia (image, audio or video) form. Textual information is stored in the form of digital documents. Digital documents are vulnerable to get copied. Copying the content without proper citation is a plagiarism. Usually in case of academic, student are bound to fall for plagiarism. Plagiarism is a serious problem in academic, publishing and research article. Number of plagiarism detection tools are available, but they follow the bag of word strategy same as that of information retrieval. But plagiarism detection is not confined to detect copy paste, but also to compare semantic associated with it. Sentences can be rearranged or replaced by synonyms conveying the same meaning that of the original. Such plagiarized sentences can easily bypass bag of word approach. Semantic analysis of the sentence helps to find such plagiarized sentences.**

In this paper, we have presented a detail survey of earlier plagiarism techniques and some of the recent techniques.

***Index Terms*- Plagiarism, semantic similarity, WordNet, copy-paste plagiarism. Natural Language Processing.**

I. INTRODUCTION

Plagiarism is the serious issue for the last two decades, it is defined in various ways as- "The theft of someone's intellectual property", "The use of someone's data, language and writing without proper acknowledgement" [1] etc. Plagiarism means copying others thoughts, ideas and concepts without giving credit to the original author, or failing to give a citation while publishing. Such dishonesty can be detected through plagiarism detection tools. Recent research found that 70% of students confess for plagiarism, with about half being guilty for cheating offense on a written assignment [11]. The person who found to be guilty shall undergo legal punishment defined by University norms [7]. Sometimes a student could fail to cite and may found to be guilty. Hence, plagiarism detection tools are needed to find and guide student to avoid such cases.

Number of plagiarism detection tools are available in the market. When we look back at early 90's, they follow the traditional approach (Vector space model) for document comparison. Each document is represented as a vector of keywords. Vectors of two documents compared using cosine similarity. Difference between documents obtained by cosine angle, as minimizes the angler maximum is the similarity. Such approach is not suitable for plagiarism detection [1, 2] as keywords can be replaced by their synonyms; sentences can be rearranged conveying same meaning. Such sentences can easily bypass a bag of words approach. Plagiarism detection is not confined to identifying copy- paste, but also analyze semantics associated with it [4]. Semantic plagiarism detection came into the picture with the rise of natural language processing technology. Researchers focused on semantic analysis and [5, 6, 9, 10] approaches proposed. Word net thesaurus is widely used to identify the semantics [6]. Here we have shown some of the earliest and the recent plagiarism detection method and found that semantic plagiarism (idea plagiarism) detection aims to provide high performance in terms of detection [3, 16].

This paper is organized into four sections. Section I, give an introduction about plagiarism and traditional method used, section II gives plagiarism scope in the field of academia, section III explain brief survey about studying approaches. Next, section IV explains architecture, methods and limitation. At last, section V concludes about the survey.

II. SCOPE OF PLAGIARISM

Plagiarism has wide scope, but here we have considered Plagiarism cases related to academia. We have classified it into two broad category viz. Plagiarism in multimedia and plagiarism in Text. Plagiarism in multimedia understood as the act of violating copyright of any multimedia file (i.e. Audio, Video, Image) whereas plagiarism in the text is when "someone is trying to steal someone else's writing and presenting as its own without proper authorization of owner". Textual plagiarism is further classified into two subcategories *a. Plagiarism in programming language* is type of plagiarism in which programmer is trying to copy whole or part of programming code without proper citation

***Mast. Dnyaneshwar Ratan Bhalearao**, Computer department, Pune Institute of Computer Technology, Dhankawadi, Pune-411043., Pune, India, Mobile:9579234394*
***Prof. Sheetal S. Sonawane**, Computer department, Pune Institute of Computer Technology,.Dhankawadi, Pune-411043.*

ISSN: 2278 – 7798

of the resource and other is *b. Natural language plagiarism*, which occurs when a researcher violates copyright of research articles written in natural language. There are several ways by which natural plagiarism can occur, such as copy-paste, by paraphrasing the content or by failing to cite the referred content. Plagiarism in Natural language further classified into two categories *a. Intrinsic plagiarism* when an author uses a single source for plagiarism and *b. Extrinsic Plagiarism*- when an author uses more than one source for plagiarism. Intrinsic plagiarism detection considers suspicious document and check whether the content is modified by another author. Extrinsic plagiarism is difficult to identify as it has a huge source of the document. Hence it is further classified into two categories *a. Monolingual and Multilingual* defined as when plagiarism occurs in same language by rearranging sentences and another is when same published work is being published in another language respectively. The figure below depicts the scope of plagiarism in academia.

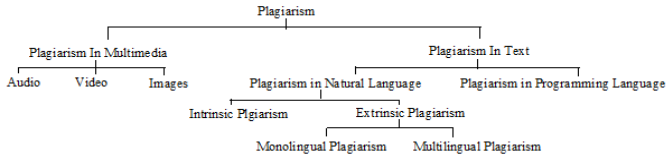

Fig 1. Scope of plagiarism in terms of academia.

III. Related Work

In this section we will be discussing about earlier tools and basic approach followed in textual plagiarism related to natural language. In the early days plagiarism detection was done by manual approach. A human manually reads the research paper and related article, detects the plagiarized content and report it to higher authority [13]. But this approach was labour intensive, the reader has to go through the content several times so, it is time consuming. This approach worked well for a limited number of document. When numbers of documents were beyond limit, human perception lacks in time and accuracy. To improve efficiency and accuracy computer based tools were developed. These tools can detect copy-paste plagiarism based on the vector space model, each document represented as the vector of words or vector of words having highest frequency count. Each vector is compared with the help of cosine similarity. Minimum the angle, maximum is the similarity. Matched documents are marked as plagiarized based on the score obtained.

We have categorized related work into two major types of tools based on strategy to avoid plagiarism viz. *A. Plagiarism Prevention and B. Plagiarism detection.*

A. Plagiarism Prevention:

This type of approach was developed to avoid to plagiarism because access is restricted to few authorized user [13]. This is pessimistic approach and plagiarizing portions of the original document cannot be identified by it. Following tool work like a copy protection system.

i) COPS (Copy Protection System) [5]*:* is a Copy Protection System for registering documents and detecting copy paste plagiarism protocol. It can identify partial or complete overlap of documents [5].

B. Plagiarism Detection:

Plagiarism detection is an optimistic approach user are allowed to access articles. It is intended to find dishonesty in the document. There are different strategies proposed based on type and scope plagiarism detection in monolingual plagiarism. We have further classified approached based on type of plagiarism detection.

i) Based on detecting of Copy-paste on word co-occurrence: These types of tools focus on word co-occurrence strategy. Vector of word is compared with vector of word from source document. Below are some of the tools which fall in this category.

a. SCAM: This presents a Stanford Copy Analysis Mechanism (SCAM) based on word occurrence frequency. It's mainly a registration server that maintains registered document which are used for copy detection. A vector of words with its frequency is used to compare with vectors in the database.

b. CHECK: A Copy detection mechanism that will eliminate unnecessary comparisons. It uses information retrieval techniques based on the fact that "a comparison between two different subjects is unnecessary. Each section, subsections and paragraph compared in detail on a sentence basis. It incorporates information retrieval techniques to create a parse tree with additional information to represent semantics of documents.

c. SNITCH: A Software Tool for Detecting Cut and Paste Plagiarism implemented with the help of Google API based on Sliding Window to Scan documents and locate candidate document.

ii) Based on Semantic analysis of the word: These types of tools focus on semantic analysis of the both documents. If they found semantically same it report a case of plagiarism. But there are two more approaches to detect semantic. One is based on word comparison of synonyms and other is based on sentence comparison.

a. Based on Semantic word Comparison:

i. *Plagiarism detection based on SCAM algorithm:* It is based on natural language. Finding the overlap of words is done by comparing sets of words that are common in the original document and document

ISSN: 2278 – 7798

testing. And the final similarity score is calculated by SCAM (Standard Copy analysis mechanism) to detect overlap of words.

b. *Based on Semantic Sentence Comparison:* Sentence comparison is done by comparing the overall sentence similarity. Following approaches work on sentence comparison.

i. *Sentence-Based Natural Language Plagiarism Detection [15]:* A tool developed for detection of natural language documents and programming assignments. Based on the concept that sentences are building blocks for communication of ideas. Proposed algorithm detects plagiarized sentence based on pairwise comparison. It is integrated with Sherlock for source code plagiarism detection.

ii. *An Improved plagiarism detection scheme based on Semantic role labeling [10]:* The basic idea is that it analyses and compare sentences based on semantic allocation of each term. SRL is used in generating argument for each sentence semantically and to analyze sentences semantically and WordNet used to extract concept or synonyms.

iii. *An Improved Semantic Plagiarism Detection Scheme Based on Chi-squared Automatic Interaction Detection [11]:* Analyses and compares text based on semantic allocation of each term of the sentence. SRL a natural language processing technique is used for extracting arguments and the important arguments are selected by using CHAID algorithm based on similarity score.

iv. *Intelligent Plagiarism Detection Mechanism using semantic technology [1]:* suggests a new strategy using semantic web. A deterministic approach for finding similar semantics of the sentences over the web. It compares user document using WordNet generally works like semantic web crawler.

iii) *Based on Syntactic Structure:* Such tools consider the syntactic structure of the sentence. A sentence tree is formed for each sentence and it is being compared with the sentence tree of the source document.

a. *Plug-In:* works on the intrinsic plagiarism detection approach by inspecting suspicious document. The main idea behind *Plag-In* is that "Author use a recognizable and distinguishable grammar to construct sentences". The Suspicious sentences can be found out by p-q gram distance of grammar trees, by using Gaussian normal distribution. The *Plag-In* algorithm compares grammar of each sentences and tries to expose suspicious ones.

IV. Architecture And Methodology

In this section we will be explaining about the methodology followed in above approaches, their advantages and limitations.

A. *COPS [5]:* architecture is designed with a goal to make it harder for being copied not to make it impossible. It is composed of two modules *first,* preprocessing that takes all registered documents and creates hash table. *Second,* Procedure for adding document whether the match is above threshold. Its Method can be depicted into three steps as *first;* Test Document first converted into canonical form. *Second,* determine sentences and generates hash keys for comparison. *Third,* find them in a hash table and if it matches is more than a threshold then report it as a violation. The system is secured by being copied but still has limitations. It lacks in detecting sentences, equations, and the abbreviation's confuses it. It cannot detect partial sentence overlap. Registration scheme can be broken by modifying words [12 and it also fail to consider individual word and take a whole sentence as one part [12].

B. *SCAM [12]:* is another copy detection system, its architecture is divided into four modules: *first,* chunking of document into words or sentences is done for document which is to be registered. *Second,* the new arrived document is chunked into same units, used for document comparison. *Third,* inverted indexing is used for each chunk (word) and document associated with that chunk. *Fourth,* if new arrived document is not registered it is compared with the each vector of the database. The main advantages are that it can detect overlapping similarity between the parts of sentences [10]. But the limitation is, it detects more false positive than COPS.

C. *CHECK [13]:* architecture composed of three modules: *first,* document registration calls document parsing module to index the document into Oracle database. *Second,* document comparison: compares with the dataset by using the Oracle database system to maintain structural characteristics of a document. *Third,* document parsing: creates internal indexing structure for each registration and comparison module.

D. *SNITCH [8]:* composed of six plagiarism detection steps *first,* read a window containing the first/next W words. *Second,* Measure the number of characters for each

ISSN: 2278 – 7798

word. *Third,* calculate the weight of the window, the average number of characters per word for the words in the window. *Four, a*ssociate this weight with this particular window for later use. *Five*, repeat the process for all such windows in the document, shifting the window forward, in the document by 1 word. *Six*, order windows in decreasing order, eliminate overlapping windows and rank all windows by weight. Select top 'N' weight windows and search on internet. *Advantage:* is that it allows variation of the size of the sliding window (W) and number of searches performed (N). But some limitation of the system are it has limitation of Google API i.e. 1000 search per day so not applicable for number of users and decreasing W leads more candidate documents but more false positive, increase (or decrease) of N will increase/ (or decrease) thoroughness

E. *Sentence based on natural language [15]:* architecture method composed of four steps preprocessing, filtering, comparison and document score. *First*, preprocessing includes changing upper to lower case, a list of words that are more common and has no meaning (e.g. A, that, the) never stored in processed form. *Second*, Filter removes the words that are repeated. *Third,* compares documents on a sentence basis. And finally, score is assigned to documents so that pair of documents can be compared. Document with high score than that of threshold and common threshold are considered to be plagiarized. Advantage of the system is that it is an integrated tool for both natural language and a programming language. There is no incident of Sherlock of missing plagiarized document. But, it lacks in identifying paraphrase, semantic plagiarism and it is a far slow approach.

F. *Plagiarism detection based on a SCAM algorithm [2]:* work on Lucene indexing. Architecture composed four modules: *first, indexing dataset:* indexing dataset is done by Apache Lucene a Java library. *Second, processing test document:* Test document converted into a token of words for searching and comparison purpose. *Third, Searching on Index:* in which index query is passed to search document from the dataset. *Fourth, evaluating similarity:* a new SCAM formula is used to calculate a similarity measure between test and source document. Advantages of the architecture is it finds overlap of words. Synonyms replacement can be identified with the use of WordNet. But, there is no standard dataset used and it cannot find the best match.

G. *Based on SRL [10]:* method is depicted in four steps *first,* text segmentation in which overall text is divided into meaningful sentences. *Second*, preprocessing which include stop word removal and stemming of the word. *Third,* Argument based label grouping (ALG) represents sentence in the form of nodes. SRL is used to extract an argument and role of each term, this extracted concept represented as graphs. *Fourth* is a Semantic term Annotation (STA) in which the synonyms and the concept that are extracted using WordNet. The advantages of the system are that it can detect copy-paste, rewording, changing word structure and modifying from active to passive and vice versa. But it has issue in selecting important argument.

H. *SRL with Chi-Square [11]:* The overall methodology can be depicted in five steps *first*, preprocessing consist of general technique such as text segmentation, stop words removal and stemming. *Second, SRL* is used to extract all the arguments. *Third,* concept extraction in which *WordNet* is used for finding synonyms and hyponyms of each term. *Next, CHAID is a statistical technique* used to select an important feature from dataset to improve plagiarism detection results and *lastly*, all the feature generated by CHAID is chosen for plagiarism detection. The main advantage is that an important argument generated by CHAID improves the result.

I. *An intelligent plagiarism detection mechanism using semantic technology:* process depicted as first a query document is decomposed into sentences, and semantic comparison can be carried out using WordNet. Finally, it searches over the World Wide Web. So it can only detect plagiarized content if words are replaced by synonyms.

A. *Plag-In [14]:* process divided in to five steps: *first,* sentences are parsed and split into sentences. *Second,* grammar trees are parsed with the help of open source tool Stanford parser, each word labeled with Penn Treebank tags and it generates grammar tree. *Third,* distance between each pair of tree is calculated and stored in a lower triangular matrix. *Four,* significant differences which are visible to the human eye in the matrix are examined through statistical methods. *Five,* it smooth the result coming from the Gaussian filter algorithm. *Advantage listed as* intrinsic approach makes it suitable for shorter document and also better in interpreting suspicious document. But the used algorism lacks in identifying suspicious shorter sentences. Fails frequently as it uses grammatical inconsistencies to find plagiarized sentences and identifies number of false positives.

V. CONCLUSION

It seems that plagiarism became a very important issue in academia. Plagiarism detection is a crucial task for every novice researcher, plagiarism detection tool help to identify plagiarized content and to guide them. We have seen the earlier and latest plagiarism detection methodologies. The

ISSN: 2278 – 7798

mentioned tools identify copy paste plagiarism in good manner, but lack in detecting complex paraphrases. So, finally, after studying the latest research paper, we found that semantic analysis can find copy paste, rewording, complex paraphrases content. The result provided by semantic analysis is near about human perception. But there is a need to enhance the semantic detecting algorithm. Hence, we suggest that semantic plagiarism detection tools will enhance the system up to human perception.

REFERENCES

[1] Agarwal, Juhi, R. H. Goudar, Pratik Kumar, Nishkarsh Sharma, Vishesh Parshav, Robin Sharma, Anubhav Srivastava, and Sreenivasa Rao. "Intelligent plagiarism detection mechanism using semantic technology: A different approach. In "*Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, pp. 779-783. IEEE, 2013.

[2] Anzelmi, Daniele, Domenico Carlone, Fabio Rizzello, Robert Thomsen, and D. M. Akbar Hussain. "Plagiarism detection based on SCAM algorithm." In *Proceedings of the International MultiConference on Engineers and Computer Scientists*, vol. 1, pp. 272-277. 2011.

[3] Alzahrani, Salha M., Naomie Salim, and Ajith Abraham. "Understanding plagiarism linguistic patterns, textual features, and detection methods." *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 42, no. 2 (2012): 133-149.

[4] Bin-Habtoor, A. S., and M. A. Zaher. "A Survey on Plagiarism Detection Systems." *International Journal of Computer Theory and Engineering* 4, no. 2 (2012): 185-188.

[5] Brin, Sergey, James Davis, and Hector Garcia-Molina. "Copy detection mechanisms for digital documents." In *ACM SIGMOD Record*, vol. 24, no. 2, pp. 398-409. ACM, 1995.

[6] Chong, Miranda, Lucia Specia, and Ruslan Mitkov. "Using natural language processing for automatic detection of plagiarism." *Proceedings of the 4th International Plagiarism Conference (IPC 2010), Newcastle, UK*. 2010.

[7] Maurer, Hermann A., Frank Kappe, and Bilal Zaka. "Plagiarism-A Survey." *J. UCS* 12, no. 8 (2006): 1050-1084.

[8] Niezgoda, Sebastian, and Thomas P. Way. "SNITCH: a software tool for detecting cut and paste plagiarism." *ACM SIGCSE Bulletin* 38, no. 1 (2006): 51-55.

[9] Osman, Ahmed Hamza, Naomie Salim, and Albaraa Abuobieda. "Survey of text plagiarism detection." *Computer Engineering and Applications Journal* 1, no. 1 (2012): 37-45.

[10] Osman, Ahmed Hamza, Naomie Salim, Mohammed Salem Binwahlan, Rihab Alteeb, and Albaraa Abuobieda. "An improved plagiarism detection scheme based on semantic role labeling." *Applied Soft Computing* 12, no. 5 (2012): 1493-1502.

[11] Osman, Ahmed Hamza, and Naomie Salim. "An improved semantic plagiarism detection scheme based on Chi-squared automatic interaction detection." In *Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on*, pp. 640-647. IEEE, 2013.

[12] Shivakumar, Narayanan, and Hector Garcia-Molina. "SCAM: A copy detection mechanism for digital documents." (1995).

[13] Si, Antonio, Hong Va Leong, and Rynson WH Lau. "Check: a document plagiarism detection system." *Proceedings of the 1997 ACM symposium on applied computing*. ACM, 1997.

[14] Tschuggnall, Michael, and Günther Specht. "Detecting Plagiarism in Text Documents through Grammar-Analysis of Authors." In *BTW*, pp. 241-259. 2013.

[15] White, Daniel R., and Mike S. Joy. "Sentence-based natural language plagiarism detection." *Journal on Educational Resources in Computing (JERIC)* 4, no. 4 (2004): 2.

[16] Yousuf, Shameem, Muzamil Ahmad, and Sheikh Nasrullah. "A review of plagiarism detection based on Lexical and Semantic Approach." In *Emerging Trends in Communication, Control, Signal Processing & Computing Applications (C2SPCA), 2013 International Conference on*, pp. 1-5. IEEE, 2013.

Dnyaneshwar Bhalerao received the B.E. in Computer Science, in 2012 from SSBT's COET, Jalgaon affiliated to North Maharashtra University. Perusing M.E. degree in Computer Engineering from Pune Institute of Computer Technology, Pune affiliated to Savitribai Phule Pune University.

ISSN: 2278 – 7798