International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. 05 Special Issue 01 | 2020

A Survey on various Efficient Plagiarism Detection Techniques

Satendra Kumar[1], Saurabh Dwivedi[2], G. Ananth Kumar[3], Vinod Kumar Chaudhary[4]

[1, 2, 3&4] *Assistant Professor, Department of Computer Science & Engineering, GL Bajaj Institute of Technology and Management, Greater Noida, UP, India*

ABSTRACT

Basically plagiarism is the practice of taking someone else's work or ideas and passing them off as one's own, which in itself is not a legal activity. And therefore must be handled with proper care. With the invention of Internet, we are able to access anything at a point of time. It is been available to all categories of people like students, employees and professionals. They can easily use the Internet as a resource for information purpose. But plagiarism is a serious issue and thus need to be eradicated in professional environment as well as in the education system. Plagiarism of report files, research papers is been done by the students without proper citing or referencing the authors of the documents. Some-times they may result to legal problems like copy infringement because of all this plagiarism detection systems come into rescue. Plagiarism can be in text, pictures, music tones etc. Main focus is to find it if it exists in real and try not to use it as it is.

Keywords: ***Tokenization, cosine similarity, SVM-Support Vector Machine, Boyer Moore algorithm, Stemming, Semantic similarity, parser, IDF-Inverse document frequency***.

1. INTRODUCTION

1.1 What is plagiarism?

First and foremost question. What is Plagiarism? Plagiarism is the practice of taking someone else's work or ideas and passing them off as one's own. There are so many types like: Copying text "as it is" without proper quotation marks and with no citation or source. In paraphrasing plagiarism, different words are paraphrased or we can say replaced by words having similar meanings. Paraphrasing without citation is also not a legal activity. Stealing idea of someone and then using it as your own. This is Idea plagiarism. Using your own past material or another student's material as a new idea without citation. This is self-plagiarism.

1.2 Types of plagiarism

1.2.1 Explicit plagiarism

When other person's words or ideas are expressed as our own, explicitly, knowing that it is obtained from other sources comes under the category of explicit plagiarism.

Example: If someone claiming that you are simply be the first one to use a certain particular idea or write it as a phrase. So it becomes quite easy for knowledgeable readers and checkers, who are acquainted and renowned about the literature in their respective fields and can easily, recognize it, that it has been taken from some other sources and not a self-made content.

1.2.2 Implicit plagiarism

When other person's words or ideas are expressed as our own, implicitly, unknowing about the fact that it is obtained from other sources comes under the category of implicit plagiarism. It happens when we do not actually claim it as our own original work or plan but still fail explicitly to point out that it is someone else's idea, there is no type of proof given to cope with this whether it is an self-created material or not as we cannot prove. This includes some part or sections which are copied from the sources with minor alterations done. It is very simple to identify such kind of thing, either as a result of their familiarity with sources within the field or as a result of the expressive style that was utilized by the first author is completely different from that found else-where within the new work, with simply modification of some particular words.

1.2.3 Unconscious plagiarism

Unconscious means the state of being unconscious, that is insensible. It is once you fail to form a note of what we have scan throughout our analysis then, have been influenced by what we have got found, later present the words and/or concepts as our own. It is our responsibility as an author to stay a record of the sources that you simply have encountered and used it in content writing. Whenever analysis is done, we need to keep a note of the required basic bibliographic data which can assist you in saving yourself from immeasurable bother in a while within the future.

1.3 Auto plagiarism

Using your own past material or another student's material as a new idea without citation. This is Auto plagiarism. It is hard to believe but it is also a type of plagiarism.

Example: We have done a research over a certain topic and then published it in some journal or officially submitted to university, later at some point of time we extended our work in the same topic and used our previous work in writing the content of new work then it will come under the category of auto plagiarism. There is a need of acknowledgement to be written while taking help from old content though written by yourself. It is simply so much similar as we tend to do with any other source, as every piece of published material should be original. Additionally, we ought to never ever submit or hand over the 'same' piece of work over more than to one teacher, university or journals.

1.4 Applications of Plagiarism detection systems

There are five reasons to use a plagiarism detection system, they are:

1. Generally people use Internet search engines to look for plagiarized material, but there is an extra edge or we can say a extra benefit as plagiarism software can offer more sources, such as large databases thousands of periodicals too that are not available online and accessible by search engines. But plagiarism checkers have access to these databases.
2. Many plagiarism detection systems highlights the content that is exact. In other words, we can see by our self what sentences or words are matched and this helps in tracing back our mistakes.
3. Most of the plagiarism detection systems also gives percentages of similarity, on the basis that how much they are similar with the online resources available. There are so many universities which use plagiarism detection systems like turn tin to check papers for plagiarism. There is a particular limit set for the students that their paper must not contain similarity above than it. If they fulfill that condition then only their thesis or paper gets submitted and thus approved.
4. It helps in paraphrasing the text as after checking it through the plagiarism detection system it gets highlighted.
5. Plagiarism detection system also is a proof to let us know like our instructors and the university that what we have written is our own created material and has not been plagiarized.

2. LITERATURE SURVEY

Literature survey is basically done to know about the domain in which we are doing research. It helps in getting more knowledge about the topics and the problems that exist in the field, which can be solved by doing research and then applying it with some solutions. For getting the content knowledge of the subject we have gone through various

Research papers and they have helped us in gaining useful information. Few of them are as follows

1. An Eccentric Approach for Paraphrase Detection Using Semantic Matching and Support Vector Machine.
2. Paraphrase Identification in Short Texts using Grammar Patterns
3. Semantic Keyword-based Text Copy Detection Method
4. Maulik: A Plagiarism Detection Tool for Hindi Documents
5. An Efficient Plagiarism Detection System using Boyer Moore Algorithm

PAPER 1: An Eccentric Approach for Paraphrase Detection Using Semantic Matching and Support Vector Machine.

Objective: Objective of this paper is to measure similarity and evaluate the paraphrase between the sentences using tokenization and SVM (Support Vector Machine) classifier.

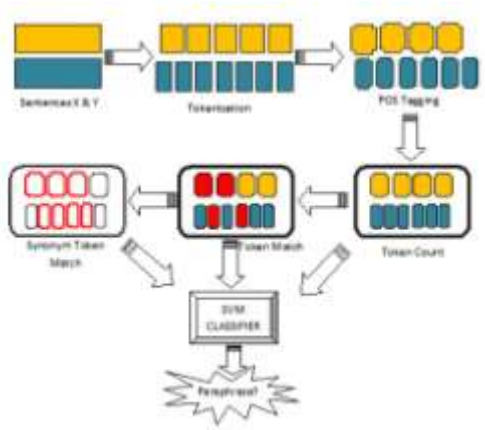

Fig-1: The Proposed Architecture of Paraphrase Detection Using Support Vector Matching

Paraphrase detection is completed by detecting word by word relation among the texts. Detection is done by means of counting the tokens in the texts and synonym of the tokens, and token matching of the sentence with the other given sentence which generates the output as +1 or -1 with help of SVM. Sentences X and Y are taken as input and then the next step is to tokenize them. After tokenization, the next step is POS tagging. Part-Of-Speech Tagger is a type of software that scans text which is written in some particular language and allot parts of speech like noun, verb, and an adjective to each word. After it tokens are counted in the token count step. Tokens are matched and also if their synonyms are obtained then they are also matched. Later they are given to Support Vector Machine. The output as +1 or -1 with help of SVM. Output as +1 means that it is plagiarized and -1 means that it is non-plagiarized [1].

Analysis results / conclusion:

Dataset used: PAN 2010. Evaluating the paraphrase between the sentences using tokenization and SVM classifier. It can also be applied on negation and antonyms.

PAPER 2: Paraphrase Identification in Short Texts using Grammar Patterns
Objective: Identification of paraphrases in the sentences by generating grammar rules / patterns and using parsing.

Approach: Here they constructed a semantic coordinate space and then computed the similarity score based on lexical databases. Parsing used to extract typed dependency and generate grammar patterns after that similarity algorithms applied and normalization done at the end.
Workflow for the approach is shown below. It takes input of 2 short length texts and then it returns a probabilistic score which indicates how much similar are the 2 different texts [2].

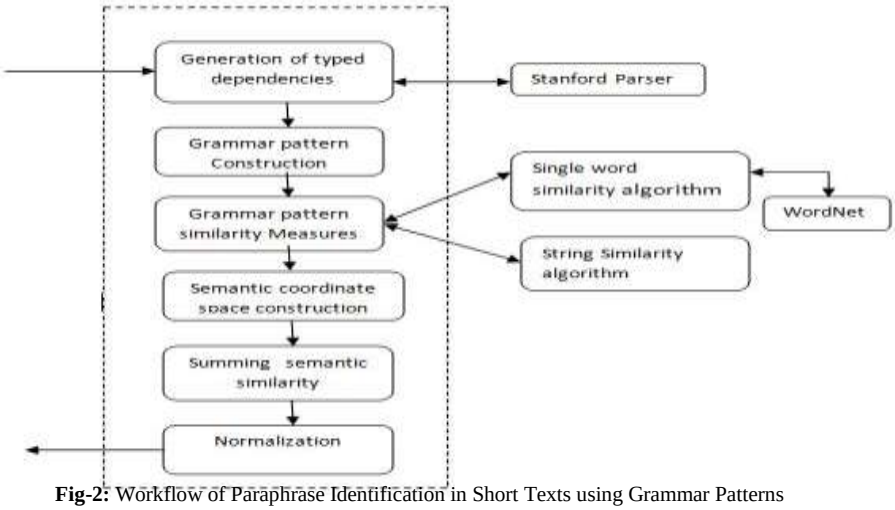

Fig-2: Workflow of Paraphrase Identification in Short Texts using Grammar Patterns

Generation of Typed dependencies: Syntactic information is obtained using the Stanford parser in case of text written in natural languages. Stanford parser is considered as a highly optimized and lexicalized dependent parser. This parser is used as it has good accuracy and it can also parse raw and non-tokenized texts.
Some training is required for applying it on domain specific texts. Stanford parser generates as output: the parser tree.
Grammar Pattern Construction: A natural language text can be rewritten in terms of a set of grammar patterns, and thus there is construction of grammar pattern.
Grammar Pattern Similarity :Grammar pattern similarity is calculated using the string similarity algorithm, more precisely the single word similarity algorithm.
Semantic Coordinate Space Construction: There is a grammar ambiguity because of large number of grammar patterns generated for each given input text. For solving this problem, they changed the grammar patterns into a unified se-mantic structural model. Grammar patterns Pi of text TA are compared against grammar patterns Pj of text TB, after that pairwise grammar pattern similarity scores are constructed in the form of a matrix representation which is known as semantic coordinate space.
Summing Semantic Similarity: After this sum of the semantic similarity is done.
Normalization: Similarity score is to be computed for various input texts, so there is a need to normalize score in the range from 0 to 1. Normalization constant is basically calculated using below equation

Normalization (NA, NB) = NA + NB / 2 * NA * NB

Where, NA and NB represents the numbers of grammatical patterns present in Text A and Text B, respectively. Similarity score of the texts must get to be multiplied by the value of normalization function so that we achieve similarity between the 2 texts.

Analysis results / conclusion: Dataset used – MSRP with the combination of various word similarity algorithms results in higher precision and performance.

PAPER 3:

Semantic Keyword-based Text Copy Detection Method

Objective: To accurately detect the plagiarism and improve the efficiency of Plagiarism detection system than when it is compared with the traditional TFIDF algorithm.

Approach: This approach consist of four components: constructing the corpus, corpus preprocessing, constructing of the feature term library and then the implementation of the copy detection algorithm.

Constructing the corpus: They selected 8 categories of documents from Chinese corpus, out of which each category has 1000 articles. The categories were mainly related to finance health, education, sports, military, tourism, culture and recruitment. Feature extraction method used in the paper is the improved TFIDF method, where the formula of IDF was modified to accurately. So that the feature extraction of each document can be finished.

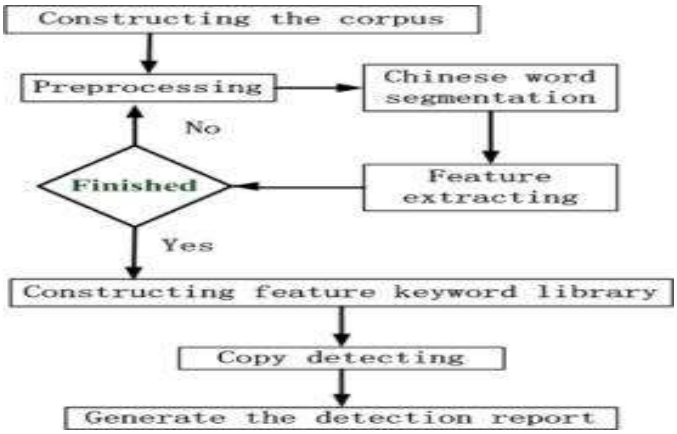

Fig-3: Workflow of Semantic Keyword-based Text Copy Detection Method

According to workflow firstly they constructed the corpus and then preprocessing is done. Chinese word segmentation is done and also feature extraction is done. It is preprocessed till all segments are preprocessed. After that feature keyword library is made. Now whether the text has been copied from somewhere is checked and later detection report is generated [3].

Analysis results / conclusion:

It proposed an improved TFIDF algorithm, which is used to accurately extract the feature words from the documents in the corpus. This algorithm is used to detect text plagiarism.

This approach is feasible, but they didn't consider the context of the sentence structure and semantics.

PAPER 4: An Efficient Plagiarism Detection System using Boyer Moore Algorithm

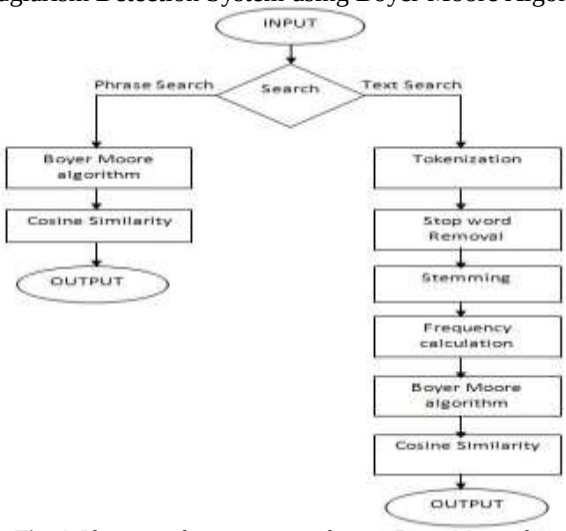

Fig- 4: Plagiarism detection approach using Boyer Moore algorithm

Here plagiarism is checked using 2 type of search: Phrase search, searching where we don't con-sider the preprocessing steps and check for exact phrase and Text search, which does not contain stop words and stemmed words. Comparison between 2 files either in .doc or .docx format or pdf format can be done by frequency calculation, removal of stop words, stemming and then search string algorithm Boyer Moore. Also efficiency is calculated in terms of similarity between the documents using cosine similarity. For calculation of efficiency of Boyer Moore algorithm in Plagiarism detection system, it uses a corpus of around 15 documents. By comparing with a variety of doc files and well as pdf files efficiency comes out to be 96.68 by considering exact match words, phrase search and modifications done to the root word.
Basically this is an offline plagiarism detection system [6].

3. CONCLUSION AND FUTURE WORK

Till now we have concluded that there are so many ways to detect whether plagiarism is there in the document or not. But the main aim is to find the best possible algorithm or way which is efficient and can calculate the plagiarism value in the respective documents. Plagiarism detection can be extended to online documents checking. Not much efficient for short substrings, there comes the existence of Brute force approach and Knuth Morris Pratt algorithms for binary strings. Future work may involve work to be done on copyright and protected pdfs and doc formats.

4. REFERENCES

[1] Sachin V. Shinde , Sangram Z. Gawali ,Devendrasingh M. Thakor "MAS a scalable framework for research evaluation by unsupervised machine learning – Hybrid plagiarism model " International Conference on Peravasive Computing (ICPC) IEEE 2015
[2] P.Vigneshvaran,E. Jayabalan, A.Vijaya Kathiravan "An Eccentric Approach for Paraphrase Detection Using Semantic Matching and Support Vector Machine" International Conference on Intelligent Computing Applications, IEEE 2014, pp 431-434.
[3] Vaishnavi V 1, Saritha M, Milton R S "Paraphrase Identification in Short Texts using Grammar Patterns" International Conference on Recent Trends in Information Technology (ICRTIT), IEEE 2013, pp 472-477.
[4] Jianjun Zhang , Xingming Sun, Jin Wang "Semantic Keyword-based Text Copy Detection
[5] Method", Advanced Science and Technology Letters 2014.
[6] Abhay Nitin Pai,Chinmay Neelmadhav Bhusari "Plagiarism Detection System" International Journal of Innovations in Engineering and Technology (IJIET), 2013
[7] Harsha Gupta,Sanjay Ojha "An Efficient Plagiarism Detection System using Boyer Moore Algorithm" International Journal of Emerging Technologies in Computational and Applied Sciences
[8] Maria Kashkur "Research into Plagiarism Cases and Plagiarism Detection Methods" Scientific Journal of Riga Technical University Computer Science 2010.
[9] Daniel louw "Dealing with Plagiarism in introductory Programming" 6th annual international conference on computer science education: innovation and technology.

BIOGRAPHIES

Satendra Kumar is working as Assistant Professor at Department of Computer Science & Engineering in *GL Bajaj Institute of Technology and Management, Greater Noida, UP, India*. He has completed his Ph.D Degree in Computer Science from GKV Haridwar. He completed his M.Tech degree from YMCA University of Science & Technology, Faridabad. He pursued his B.Tech degree from MJP, Rohilkhand University Bareilly. He has published more than 20 research papers in repute journals and conferences.

Saurabh Dwivedi is working as Assistant Professor at Department of CSE in GLBITM Greater Noida. He has completed his M.tech Degree in CSE from Galgotias University.

	G. Ananth Kumar is working as Assistant Professor at Department of CSE in GLBITM Greater Noida. He has completed his M.tech Degree in IT from JNTUK (AP).
	Vinod kumar chaudhary is working as Assistant Professor at Department of CSE in GLBITM Greater Noida. He has completed his M.tech Degree in CE from YMCA.