An Integrated Approach for Plagiarism Detection System

Shilpa[1], Mr. Manoj Challa[2]
[1,]CMRIT, BANGALORE,
[2] Assoc Professor, Dept of CSE,CMRIT, Bangalore,

ABSTRACT:
Nowadays as Internet is becoming the primary media for information access and nearly every information is available in the Internet. Therefore, it becomes easier to use another author's contents from the Internet without giving proper reference. Others ideas, solutions or expressions are representing as one's own original work is known as plagiarism. In this paper, we propose a framework which works by integrating several analytical procedures. Scholarly documents under investigation are segmented into logical tree-structured representation using a procedure called DSEGMENT. Statistical methods are utilized to assign numerical weights to structural components under a technique called C-WEIGHT. The top weighted components from the structural components are fed into plagiarism detection technique called Optimized Semantic Role Labeler (OSRL).This technique analyses and compares text based on the semantic allocation for each term inside the sentence. The absolute hash function method is used to map large data sets of variable length to smaller data of fixed length by generating hash key .In terms of Recall and Precision, this method outperforms the results with the existing plagiarism detection methods.

KEYWORDS: *Plagiarism detection, Scientific Publication, Semantic Argument, Semantic Role, Semantic Similarity, Natural language processing, hash map.*

I. INTRODUCTION

P lagiarism is defined as the appropriation or imitation of the methods, ideas and views of another writer and representation of them as one's original work[1]. Plagiarism in scientific publications has increased and one day you will see your published work is used in another publication yet without proper references and citation. Scientific publications in the same field normally share the same general information. Besides, each publication should express a particular message that contributes to that field. Diverse contributions can be made in diverse areas; for example, solving new problems, giving suggestions to existing problems, experimenting new methods, comparing current methods, giving enhancement to results. Such these contributions of others are considered their ideas and should be acknowledged when reported in a further research. The big challenge is to provide plagiarism detecting with proper method in order to improve the time required to check the results and percentage of finding results. Several plagiarism detection tools use character or string matching method to detect the plagiarized content. Most of the current software's and techniques are less effective in detecting a plagiarized text because these tools tend to compare the suspected text with original text using characters matching, some with chunks while others by words example MOSS[3],jplag[4]. This leads to rigorous search which takes a long time in the matching process. The matching algorithms are working depending on the text lexical structure rather than semantic structure. Therefore it becomes difficult to detect the text paraphrased semantically. One of the goals of this study is to propose new semantic techniques for plagiarism detection based on optimized Semantic Role Labeler. The proposed method does not analyze the content of a text document as text syntax only, but also captures the underlying semantic meaning in terms of the relationships among its terms.

In this paper we propose a new plagiarism detection technique based on analytical procedures and Optimized Semantic Role Labeler. In analytical procedures we highlight the similar sections by giving the statistical weights. Optimized Semantic Role Labeler can help in detecting sophisticated obfuscation such as copy paste, anonyms, renaming, paraphrasing, splitting, merging, and changing the sentence from active to passive voice and vice versa in the highlighted document. The absolute hash function is used to map large data sets of variable length to smaller data of fixed length by generating hash key.

By using the absolute hash function we can reduce the retrieving time in the list. The remaining section of the paper is structured as follows; Section II discusses the related literature survey on plagiarism detection. Section III discusses the framework for plagiarism detection using OSRL. Section IV explores the experimental analysis. In section V, we discuss the results and the potential market of OSRL, and we draw a conclusion and future of this work in the last section.

II. RELATED WORK

The above section gives a overview of the existing plagiarism methods. Many tools works on finger print based method, String matching method. Fingerprint method [5] [9] works by creating fingerprint for each document in the collection. It gives the detailed information about the number of terms per line, number of keywords, and number of unique terms. Mostly rely on the use of K- grams (Manuel et al. 2006) because the process of fingerprinting divides the document into grams of certain length k.Different plagiarism tools have been surveyed in multiple works. Many online solutions are established today; for example, CrossCheck,Turnitin,SafeAssign,EVE2, WCopyFind ,Viber, Scriptum, PlagiarismDetect, SCAM ,CHECK , PPChecker , SNITCH , Ferret and others. Several studies have reported that using plagiarism checker tools in academia is effective in reducing the problem, and discouraging the students to commit plagiarism . Besides, anti-plagiarism tools help to educate students and authors in different disciplines about plagiarism .

JPlag [4], Sherlock are the examples of String Matching- based algorithm. Many String Matching algorithm works based on tokenization, where each file is replaced by predefined tokens. One of the problem of JPlag[4] is that files must be parse to be included in the comparison for plagiarism because of this some similar files that are not parsed to be missed. Unlike JPlag, in Sherlock files do not have to parse to be included in the comparison and also displays quick visualization. The time complexity and space complexity is more in the above method and also not checking at sentence level. DANIEL R WHITE and MIKE S JOY[6] proposed a method by comparing suspected documents at sentence level there by detects paraphrasing, reordering, merging ,splitting but fails to analyze semantically and also speed is less compared to other software like CopyCatch.

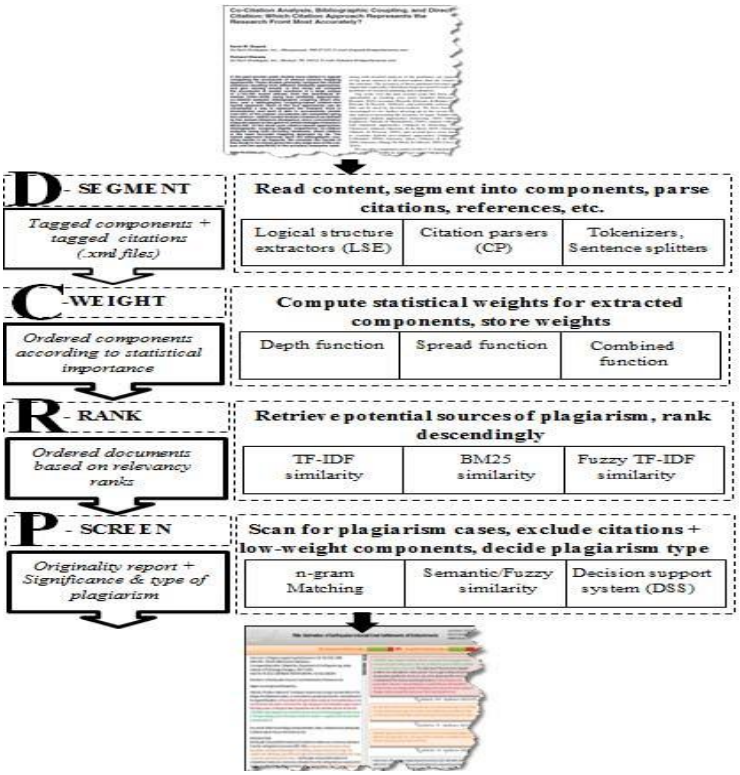

Figure 1: Existing plagiarism detection framework Iplag

Iplag[7] gives a accurate judgment about type of plagiarism. It works by combining several analytical procedures. Documents under investigation are segmented into D segment [7].Then weights are given to D-segment called as C-weight. Relevance Ranking(R-Ranking) and plagiarism Screening approach [7] (P-screen) is used. The architecture for Iplag which is useful for existing plagiarism detection systems is shown in

figure1.This Iplag acts as a framework and can be applied to any of the existing plagiarism detection technique.Plagiarism Detection Scheme Based on Semantic Role Labeling [2] analyses and compares text based on the semantic allocation for each term inside the sentence. It identifies and label arguments in a text. The idea behind SRL is that the sentence level semantic analysis of text determines the object and subject of a text. For calculating the semantic labeling, to retrieve semantic arguments and labels in large document will take large amount of time and space. The architecture as shown below

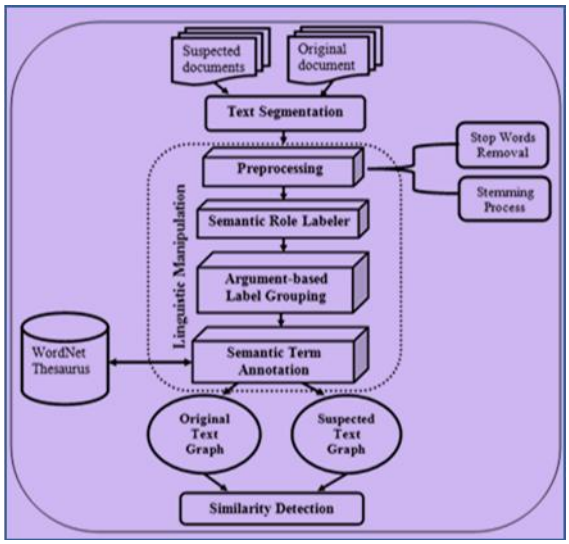

Figure 2: Structure of Semantic Role Labeler

III. FRAMEWORK FOR PLAGIARISM DETECTION USING OSRL

This paper proposes a framework that is very accurate and reliable in reporting plagiarism in scientific publications. Checking plagiarism manually requires an Expert. An Expert must have idea about the contributions in different parts of paper like discussions, conclusions and any suspected methods, techniques, ideas that might be taken from somewhere else. The experts may not pay attention to every statement but only sentences that convey original ideas and ignores important parts. The OSRL is used for sentence level semantic analysis of the text that determines the object and subject of a text. It relays on the characterization of events such as determining "when", "where ", how: who" did,"what" to "whom". The predicate of a clause (normally a verb) establishes "what "took place and other parts of the sentence express the other arguments of the sentence. The main goal of OSRL is to identify what semantic relation holds among a predicate and its associate participants with relations drawn from a predefined list of possible semantic roles for that predicate.

3.1.Document Segmentation [D-segment]

The framework begins by splitting the scientific publications into several meaningful components [7] i.e.

- document->sections->paragraphs
- document->paragraphs->sentences
- document->topics->paragraphs

In this framework, the logical structural organizations of scientific publications were used in which it includes most of the following categories: Title, Owner, Abstract, Introduction, Literature survey (related previous works, etc.), Evaluation, Acknowledgements, and References. These categories are called generic classes. Further, generic classes involve different structural components such as head titles, paragraphs, tables, tablecaptions, equations, figure captions; etc, Structural components are further segmented into sentences.

3.2.Component Weighting: C-WEIGHT

By assigning numerical weights to structural component .we can show how important that component is to the article. Depth and Spread functions are statistical measures and are useful for detecting results. Spread of a term t defines the number of structural components that contain the term. Depth of a term t refers to the frequency of the term in a class (unlike normal term frequency which considers the number of the term's occurrences in the whole document).The weight w of a structural component c in a generic class Gc can be obtained by clubbing Spread and Depth ,that can be expressed as follows:

$$w(c) = \sum_{t \in c} spread(t) \times \frac{tf_{t,G_c}}{\max t f_{G_c}} \quad (1)$$

Where spread (t) is the number of components that has t, tft, Gc is the frequency of *t* in the generic class *Gc* which has *t*, and $\max tf_{Gc}$ is the maximum frequency occurs in that class. Select the top weighted components that are having more weight.

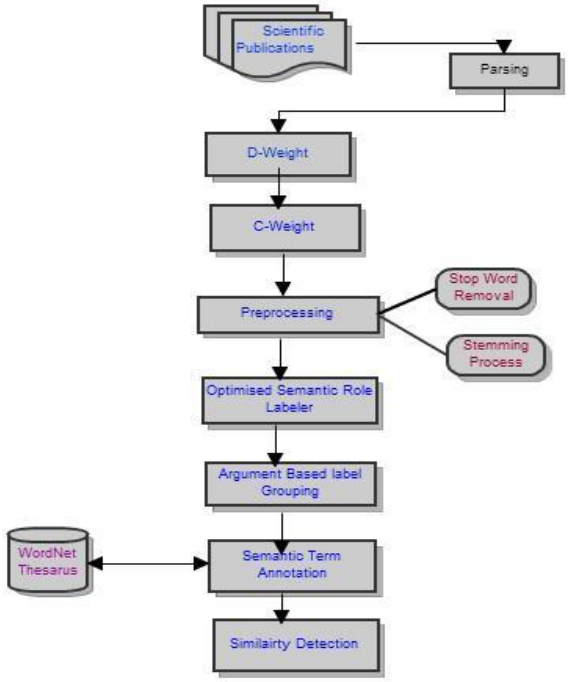

Figure 3 .Presents a Plagiarism Detection Framework procedure

3.3.OSRL for plagiarism detection

In this method, we take top weighted components of the suspected documents for pre-processing steps .i.e. stop words removal and stemming. Then, OSRL was used to transform the sentences into arguments based on location for each term in the sentences. The verbs play an important role in every sentence for processing, and the analysis of the sentences. All the arguments extracted from the text were grouped in the nodes according to the argument type. Each group was named by the argument name. This step is called Argument Label Group (ALG). Then, we extracted all the concepts for each term in the argument groups using WordNet thesaurus. This step is called Semantic Term Annotation (STA).

3.4.Similarity Detection

Create a hash values for the labels created from the OSRL. Similarity Detection between original and suspected document is done by first extracting the hash values of the suspected document and then check that with hash values of original documents. If two is matching then start finding similarity of labels with respect to hash values which are matched.

Plagiarist can be shown through the following example:James gave Rohan the parcel (original sentence).
The parcel was given to Rohan by James (suspected sentence)

By using OSRL the produced arguments are:
James gave Rohan the parcel
Output:

OSRL *Charnaik*

James	giver [A0]	(S1 | S |NP (NNP James)|
gave	V:give	|VP (VBD gave)
Roha n	entity given to [A2]	(NP (NNP Rohan | |
the	thing given[A1]	(NP (DT the |
parcel		(NN parcel | | |
.		(. .)))

Figure 4. Analysis for original sentence using OSRL

The parcel was given to Rohan by James
Output:

	OSRL	*Charnaik*
The	thing given [A1]	(S1 | S (NP (DT The |
Parcel		(NN parcel))
Was		(VP (AUX was)
Given	V:giv e	(VP (VBN given)
To	entity given to [A2]	(PP (T0 to)
Rohan		(NP (NNP Rohan)))
By	giver[A0]	(PP (IN by)
James		(NP (NNP James)))))
.		(. .)))

Figure 5. Analysis for suspected sentence using OSRL

Fig 4 and 5 explains how the suspected sentence analyzed using OSRL. The structure of two sentences may different if synonyms and antonyms or active and passive are used. But they might be same through semantically. The OSRL captures the arguments for a sentence in spite of changing the places for the labels inside the sentences. This method of analyzing the sentence supports our proposed method in detecting plagiarism if comparison is applied based on the arguments of the sentence using OSRL.

3.5.Experimental Design and Dataset

The experiments were performed on 100 suspected documents, each plagiarized from one or more original documents. In this point, Original and tokenized suspected documents were analyzed by sentence-based similarity. Sentences in suspected documents were compared with each sentence in the candidate documents according to the arguments of the sentences. We not only detect the arrangement similarity between sentences, but also possible semantic similarity between two sentences. Similarity detection was conducted by comparing the original Topic Signature terms and suspected Topic Signature terms. If the two terms were found to be identical, we went directly to the argument label groups that contained these terms, and then determined the label group where they belonged, thus determining the possible sentences that may be plagiarized. This step compared the arguments of possible sentences that had been plagiarized with the corresponding arguments in original sentences. The argument label group gave way to the main arguments and each argument inside the group quickly take to the possible plagiarized sentence. Some parameter play a crucial role in the similarity calculation, such as the number of matched arguments and number of arguments which exist in the sentences. The first variable determines the similar arguments between the suspected document and original document

while the second variable determines the argument that does not exist in the sentences. The similarity between the arguments of the suspected document and original document was calculated according to Jaccard coefficient measure [8] which is well-known as famous similarity measure between two sets. Jaccard coefficient [8] defined as a following equation:

$$Similarity\ C_i(ArgS_j, ArgS_k) = \frac{C(ArgS_j) \cap C(ArgS_k)}{C(ArgS_j) \cup C(ArgS_k)} \quad (2)$$

Where,

C(ArgS_j)=concepts of the argument sentence in the suspected document; C_i(ArgS_k) =concepts of the argument sentence in the original document;

We then calculated the similarity between the suspected document and original document based on the following equation:

TABLE 1.
EVALUATION MEASURE OF THE PROPOSED METHOD

Number of documents	Recall	Precision	F-measure
1000	0.828237	0.655606	0.761109

$$\text{Total Similarity (Doc1, Doc2)} = \sum_{i-1}^{l} \sum_{\substack{j-1 \\ k-1}}^{m \\ n} Sim C_i(ArgS_j, ArgS_k) \quad (3)$$

Where,

SimC_i (*Arg*S_j, *Arg*S_k) is similarity between Arguments sentence j in suspected document containing concept i Arguments sentence k in original document containing concept i , l = no. of concepts, m= no. of Arguments sentence in suspected document, n = no. of Arguments sentence in original document.

IV. RESULTS AND DISCUSSION

Our technique was tested according to the group of 100 documents .Those suspected documents were plagiarized with different ways of plagiarism such as simple copy and paste, changing some terms with their corresponding synonyms, and modifying the structure of the sentences (paraphrasing). We provided three general testing parameters that are commonly used in plagiarism detection as following

$$\text{Recall} = \frac{Number\ of\ Detected\ Arguments}{Total\ Number\ of\ Arguments} \quad (4)$$

$$\text{Precision} = \frac{Number\ of\ Plagiarized\ Arguments}{Number\ of\ Detected\ Arguments} \quad (5)$$

$$\text{F-Measure} = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (6)$$

The proposed method is evaluated and compared with Latent Semantic Analysis technique. The results are shown below

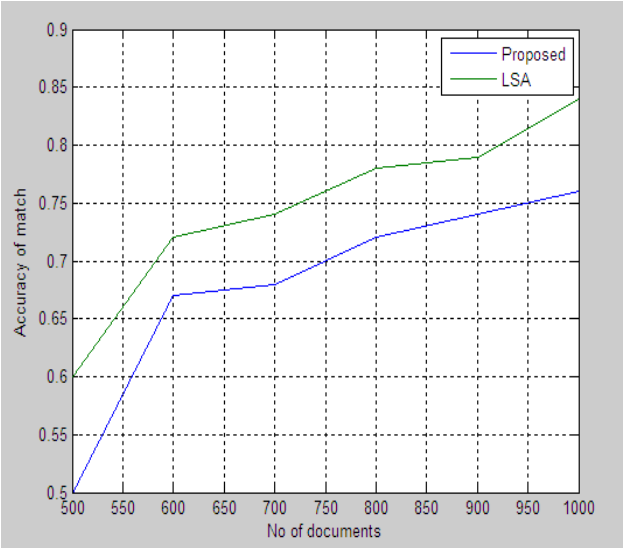

Figure 6. Comparison of two techniques in terms of accuracy.

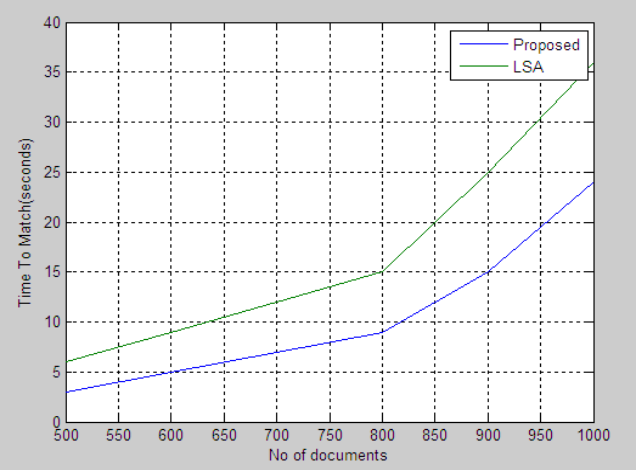

Figure 7. Comparison of two techniques in terms of time required for plagiarism detection.

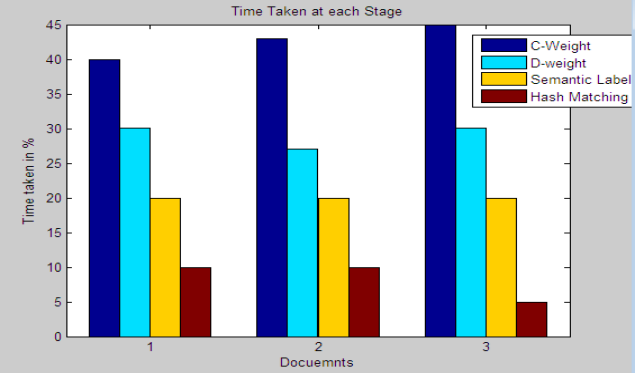

Figure 8. Time taken in each modules present in OSRL technique.

Fig. 6 and 7 demonstrates the comparison between OSRL and Latent Semantic Analysis in terms of accuracy and time required for Plagiarism Detection respectively. Fig 8 demonstrates the time required in each module which are present in OSRL technique. We found that all the scores that were obtained by our proposed method have good results than other method. The results from the comparison show that the proposed method achieved better results in terms of accuracy and time required for Detection.

V. CONCLUSIONS AND FUTURE WORK

Optimized Semantic role labeling can be used for plagiarism detection by extracting argument of sentences and comparing the arguments. Tests were carried out against 100 dataset for plagiarism detection. The proposed methods were found to achieve better performance compared to Latent Semantic Analysis [9]. Our future work is to improve its efficiency and time complexity by introducing Hadoop Map Reduce Technique. As of now our detection technique works for mono lingual documents. We can extend it for multi lingual documents.

REFERENCES

[1] www.academia.edu/689297/Source_Code_Plagiarism_-a_Student_Perspective

[2] Naomie Salim, Ahmed Hamza Osman, "Plagiarism Detection Scheme Based on Semantic Role Labeling", International conference march 2012.

[3] Http://theory.stanford.edu/aiken/moss/, 2005.

[4] Lutz Prechelt , Guido Malpohl ,"JPlag: Finding plagiarisms among a set of programs" March , 2000 , http://wwwipd.ira.uka.de/EIR/

[5] http://www.dcs.warwick.ac.uk/report/pdfs/cs-rr-440.pdf

[6] DANIEL R. WHITE and MIKE S. JOY," Sentence-Based Natural Language Plagiarism Detection",IEEE ACM-TRANSACTION August 23, 2005

[7] Salha Alzahrani , Naomie Salim , Ajith Abraham," iPlag: Intelligent Plagiarism Reasoner in Scientific Publications",International Conference,2011.

[8] P. Jaccard, "Etude comparative de la distribution florale dans une portion des Alpes et des Jura," Bulletin de la Society Vaudoise des Sciences Naturelles, vol. 37, pp. 547–579, 1901.

[9] Cosma, Georgina and Joy, Mike. (2012) An approach to source-code plagiarism detection and investigation using latent semantic analysis. IEEE Transactions on Computers, Vol.61 (No.3). pp. 379-394. ISSN 0018-9340

Mrs. Shilpa is presently doing Master of Technology in Computer science and Engineering at CMR Institute of Technology, Bangalore, Karnataka. She received her Bachelor of Engineering degree in Information Science and Engineering from Kalpatharu Institute of Technology, Tiptur, Karnataka in the year 2007.

Mr. Manoj Challa is pursuing Ph.D(CSE) in S.V.University, Tirupati, India. He completed his M.E(CSE) from Hindustan College of Engineering, Tamil Nadu in 2003. He is presently working as Associate Professor, CMR Institute of Technology, Bangalore. He presented nearly 18 papers in national and international conferences. His research areas include Artificial intelligence and computer networks.