Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A Novel Approach for Plagiarism Detection Using Semantic Web

Reenu Dutta[1], Dr. Lalit Sen Sharma[2]
[1,2] Department of Computer Science &IT, University of Jammu J&k, India

Abstract: ***Internet has equipped people with a great amount of information at hand. Accessing any information on the web is just a matter of a single click. But on the other hand it has lead to a serious threat of plagiarism, as the increased volume of information makes it easier for a person to get any information for copying about a specific topic. Plagiarism is a practice where the work of some author is wrongfully copied by someone and presented as their own. There are number of techniques available for plagiarism detection but their main focus is on keywords matching which proves to be inefficient for detecting modified plagiarism. This paper presents a different approach of detecting plagiarism using semantic web and ontology. In this study two documents are compared against plagiarism where initially pre-processing is performed on both documents. Then the ontologies of the documents are created using protégé editor and finally the individuals of the ontologies are compared using WordNet which gives the degree of plagiarism between the two documents. The experimental results show that the semantic web technologies can be used effectively to check modified plagiarism.***
Keywords: ***Internet, plagiarism, semantic web, ontology, modified plagiarism, WordNet, ontologies***

I. INTRODUCTION

Plagiarism is the act of imitating someone's information without consent and naming that as their own. Plagiarism has become a very common exercise as a result of the availability of vast amount of electronic data present online for open access. Data on internet is growing in size each day making it easier to get any information in few seconds only thereby making plagiarism detection very difficult. Diagnosis of plagiarism manually is next to impossible. Plagiarism is considered as an educational dishonesty and influence the excellence of research. There are number of techniques available but are not capable of discovering modified plagiarism.

The exact copy plagiarism [1], where the contents of the original document are simply lifted from the source document and copied into the plagiarized document without making any changes is easier to detect by simply keywords matching. And there are many techniques available for detection of this type of plagiarism which shows good results. There is other type of plagiarism, the modified copy plagiarism [1], where the plagiarist performs certain modifications on the content before copying it like rearrangement of words or replacement of words with synonyms. This type of plagiarism is difficult to detect and generally circumvent the plagiarism detection. Number of techniques have been developed but proved inadequate to detect plagiarism of this type. Therefore we tried to find out a technique which would detect the modified plagiarism effectively.

Semantic web associates the semantic of a document along with it which is represented by formal ontologies, providing shared conceptualization of specific domain [Gruber 1993] [1]. This paper presents a plagiarism detection technique based on semantic web technologies specifically ontology. The structure of ontology is used effectively to store the knowledge or concepts contained in a document. In our work we have created ontologies of the suspected document and the original document using protégé editor. Protégé is a free, open source ontology editor and knowledge base framework that provides a suite of tools to construct domain models and knowledge-based applications with ontologies [2].

II. AIMS AND OBJECTIVES

The aim of the work under study is to find a suitable approach to detect plagiarism which has been carried out by performing modifications on the document either by replacing the words in a document by their synonyms or modifying document by changing the position of words in the document keeping the idea of the document unaltered.

III. RELATED WORK

Detection of plagiarism is vital for educational institutions in order to eliminate the unethical practice of plagiarism. Many researchers are carrying their respective study in the domain some of them are as under:

Ion Smeureanu et. al., [2] introduced a source code plagiarism detection system using protégé editor. They have showed that

ontologies can be used in detecting source code plagiarism. They created the ontologies of the respective source codes in protégé using the web ontology language OWL. Finally they applied SPARQL on both the ontologies to extract the needed information from the ontologies.

Eman Salih Al-Shamery et. al., [3] have described an approach to detect semantic plagiarism using WordNet dictionary. They have worked on the location of the words in the document considering that if the word is replaced by the synonym but the position is not altered then semantic plagiarism is present.

Juhi Agarwal et al., [4] have proposed architecture and algorithm that can detect the plagiarism using words and their meanings of two documents through matching the keywords of the documents.

Deepika et. al., [5] introduced a technique to detect text plagiarism using domain ontologies created offline. They used the domain ontology to extract the relevant concepts and relation of the document and used WordNet to detect the modified contents of the document.

Osman et al., [6] proposed to detect plagiarism using graph based method where the document was converted into a graph and the nodes and edges were considered. The nodes contained the sentences linked to each other through edges representing the attributes of the sentences. The nodes were connected to each other on the basis of their order in the document and all the nodes were connected to the root node "topic signature". The content of the nodes of the two documents were compared.

Tao Chi et al., [7] performed text similarity calculation based on ontology model where the data was taken from student's answers and standard answers which were first converted into ontology and the individuals of the ontology were compared using hybrid word similarity calculation method.

Taiseer Abdalla Elfadil Eisa et al., [8] have analyzed the existing plagiarism techniques and had compared their performance. They concluded that the existing plagiarism detection techniques were not efficient in detecting plagiarism.

IV. IMPLEMENTATION METHODOLOGY

The proposed method compares the content of two documents against plagiarism using WordNet dictionary. The method mainly comprises of three steps including:

A. Document's pre-processing.
B. Creation of ontologies of the documents.
C. Calculating the degree of plagiarism.

A. Document's Pre-Processing:

The pre-processing consists of the steps shown in the figure 1 below:

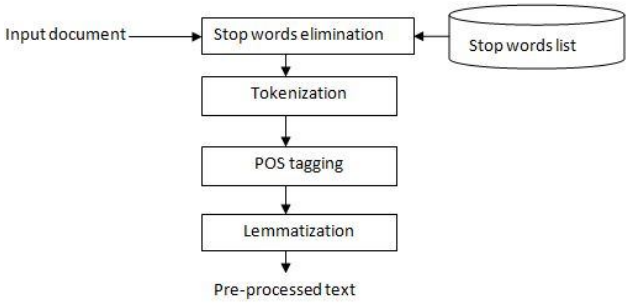

fig. 1 Steps involved in data pre-processing

1) *Stop Words Elimination:* A document incorporates several words which contain no significant information about the semantic of the document although are required for the document to be syntactically legitimate and readable. These words are frequently duplicated in the document however their elimination does not alter the concepts contained in the document. In this step all the stop words are removed from the document. A list of stop words has been used .The words of the document are compared with the contents of the list. If the word is contained in the list it is removed from the document.

2) *For example: the two texts are given*

a) *First:* "Totalitarianism is a governmental structure where the state makes no ceiling of its power and seeks to conduct every aspect of the public and private life at any time possible. A peculiar form of totalitarian government is a detailed ideology a

Volume 5 Issue VI, June 2017
ISSN: 2321-9653

collection of ideas that provides meaning and direction to the whole society".

b) *Second:* "Totalitarianism is a political system in which state recognizes no limits of its authority and strives to regulate every aspects of public and private life whenever feasible. A distinctive feature of totalitarian government is an elaborate ideology, a set of ideas that gives meaning and direction to the whole society".

It can be seen that the two texts are similar in meaning but the words are not same. Words of first text are replaced with their synonyms in the second text keeping the idea same. We would perform plagiarism detection on the two texts. Figure 2 shows the two texts after stop words elimination.

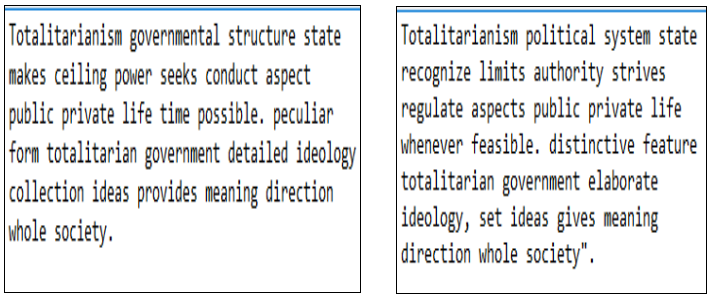

Fig. 2 Snapshots of the two texts after stop words elimination.

TABLE I
STOP WORDS LIST [9]

a	be	for	into	no	should	us
about	because	from	is	nor	since	was
above	been	get	it	not	so	want
across	but	had	its	of	some	we
after	by	has	just	off	than	were
all	can	have	least	often	that	what
almost	cannot	he	let	on	the	when
also	could	her	like	only	their	where
am	dear	hers	likely	or	them	which
among	did	him	may	other	then	while
an	do	his	me	our	there	who
and	does	how	might	own	these	whom
any	either	however	most	rather	they	why
are	else	i	must	said	this	will
as	ever	if	my	say	to	with
at	every	in	neither	says	too	would
yet	you	your				

3) *Tokenization*: Now we are left with only with the central words of the document containing the prime concepts of the document. In this step the document is parsed sentence wise and chopped up into individual words called as tokens. The tokens are used for further pre-processing. Tokenization is performed so that the system can work on individual words of the document separately.

4) *Parts of Speech Tagging:* In this step each individual is marked to its respective part of speech on the basis of its usage in the sentence. We have regarded the various parts of speech of the words as the concepts of the ontology and represented their relations. Each word used in the document is a noun, verb, adverb, adjective, etc. we have used Stanford CoreNLP for the parts of speech tagging of the individuals of the document.

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

5) *Lemmatization:* The plagiarist generally changes the form of the word in case of paraphrased type of plagiarism. Therefore we have removed the inflectional endings of the words and obtained its root word. For example, the words 'organized', 'organizing' and 'organizes' all results to word 'organize' This step was also performed using the Stanford Core NLP. The results of tokenization, pos tagging and lemmatization are shown in figure 3.

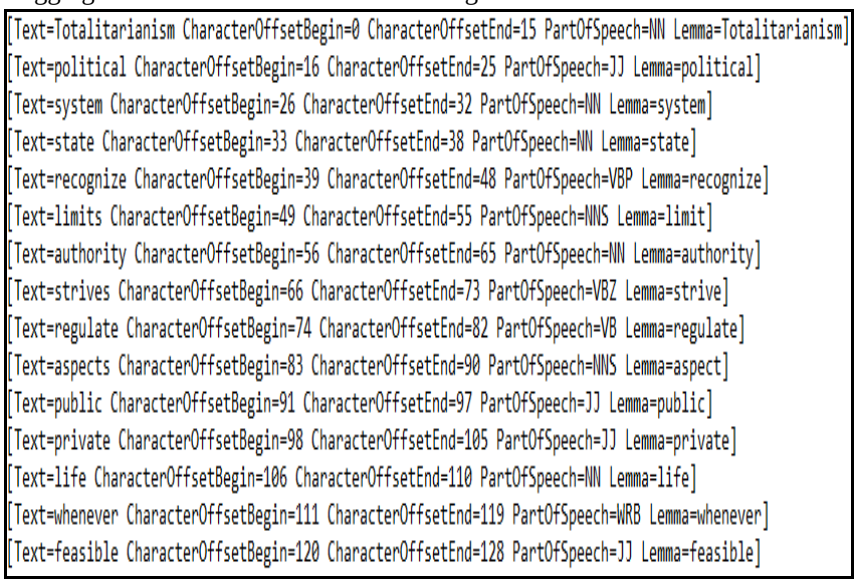

Fig. 3 Snapshot of POS tagging, tokenization and lemmatization.

B. Creation of ontology

In this step the documents are converted into ontologies. We have used protégé editor for creating the ontologies of the documents. For calculating the plagiarism between two documents two parameters need to be dealt with. First is the degree of similarity between the words contained in a document and the other is the semantic arrangement of the document. By comparing the words using WordNet the semantic similarity between the concepts of the two documents is obtained and using ontology the semantic arrangement of the documents is analyzed. The parts of speech of the document like verb, adverb, noun, adjective, etc. are made concepts in the ontology. The different words are made the individuals of the ontology according to the results of the POS tagging. The individuals are connected according to the relations they have in the document.

Table II
Shows the individuals of the two ontologies

	Noun	Adjective	Verb	Adverb
Ontology First	totalitarianism, structure, state, ceiling, power, aspect, life, time, form, government, seek, ideology, collection, idea, direction, society	public, possible, totalitarian, whole, detailed, peculiar	mean, make, seek, conduct	null
Ontology second	totalitarianism, system, state, authority, limit, aspect, life, feature, idea, government, ideology, direction, society	political, public, private, feasible, distinctive, totalitarian, elaborate, whole	recognize, strive, regulate, set, mean, give	whenever

C. Calculating the degree of plagiarism

WordNet has been used for obtaining the similarity between the words contained in the two documents and the corresponding individuals of the two ontologies. For example, the similarity between the word 'power' and 'authority' is obtained which comes out to be 0.8. We then calculate the arithmetic mean of the maximum similarities obtained above. This mean value gives us the similarity of the semantics and structures of the two documents. The similarity of the two documents comes between 0 and 1.79. Figure 3 shows the results obtained for words comparison in WordNet. When the two words are exactly similar the similarity comes to be 1.79. When the two words are synonyms of each other, the similarity is greater than the threshold (ω).

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

```
totalitarianism -  totalitarianism = 1.79769313486231570E308
totalitarianism -  structure = 0.26666666666666666
totalitarianism -  state = 0.23529411764705882
totalitarianism -  ceiling = 0.2222222222222222
totalitarianism -  power = 0.42857142857142855
totalitarianism -  aspect = 0.3157894736842105
totalitarianism -  life = 0.4
totalitarianism -  time = 0.375
totalitarianism -  form = 0.375
totalitarianism -  government = 0.625
totalitarianism -  seek = 0.375
totalitarianism -  ideology = 0.375
totalitarianism -  collection = 0.6153846153846154
totalitarianism -  idea = 0.4
totalitarianism -  direction = 0.25
totalitarianism -  society = 0.7142857142857143
system -  totalitarianism = 0.25
system -  structure = 0.8
system -  state = 0.47058823529411764
system -  ceiling = 0.6666666666666666
system -  power = 0.2857142857142857
system -  aspect = 0.21052631578947367
system -  life = 0.26666666666666666
system -  time = 0.25
system -  form = 0.25
system -  government = 0.25
system -  seek = 0.25
system -  ideology = 0.25
system -  collection = 0.3076923076923077
system -  idea = 0.26666666666666666
system -  direction = 0.5
system -  society = 0.2857142857142857
state -  totalitarianism = 0.23529411764705882
state -  structure = 0.5
state -  state = 1.79769313486231570E308
```

Fig. 4 Snapshot of word-word similarity in WordNet

V. RESULTS AND DISCUSSIONS

For the plagiarism detection, two texts have been taken the second text is the paraphrased form of the first. Some words in first text are replaced with their synonyms but the idea is kept same. WordNet comparisons are performed and two parameters are evaluated. First the values of similarity between the words contained in the whole document and second the similarity in the structure of the two documents. If both the similarities are greater than a pre-defined threshold then the suspected document is considered to be plagiarized. The two similarities are obtained by calculating the arithmetic mean of the maximum similarities between words and corresponding individuals. Table 3 shows the similarity results obtained for the two texts using the method.

TABLE III
SIMILARITY STATISTICS

Similarity parameter	Value of similarity
Word-Word	1.06
Noun-Noun	1.39
Adjective-Adjective	0.70
Verb-Verb	0.60
Adverb-Adverb	0

It has been observed that the similarity between the words in the two texts is greater than 1, showing that most of the words in the suspected text are directly copied. The similarity between the nouns, adjectives, verbs of the two texts is also very high showing that

plagiarism in structure of the sentence is also present. Analyzing the results of table 1 it can be concluded that the two documents are plagiarized.

In our study while evaluating the efficiency of the plagiarism detection system, we observed some result's value to be true positive (TP), False positive (FP), True negative and false negative. Table 4 shows the values of accuracy, precision, recall and F-measure of the system.

TABLE IV
PERFORMANCE ANALYSIS

parameter	value
accuracy	0.72
precision	0.85
recall	0.75
F-measure	0.75

The worthiness of the system is that it is efficient in detecting the modified form of plagiarism as it compares the semantic meaning of the two documents. But its efficiency depends upon the POS tagger and the WordNet dictionary. In case if the word is not present in the dictionary the similarity obtained for it is 0 which may affect the results.

Plagiarism detection can also be performed by creating the ontologies automatically using automatic ontology creating tools and then mapping the two ontologies using automatic ontology mappers [10]. But the efficiency of plagiarism detection in that case is not good and depends completely on the tools that have been used.

VI. CONCLUSION

As of now there were number of techniques available for plagiarism detection, but their efficiency decreases as certain modifications were performed on the document like rearrangement of text or replacement of words with their synonyms. As their main focus is on keywords matching the modified type of plagiarism can be bypassed. So there is a need of a strong technique which can detect the modified type of plagiarism efficiently. In the study we have used semantic web technology to detect modified plagiarism. On the basis of the results obtained after pre-processing the documents are converted into an ontology using protégé editor. The concepts of the two ontologies and the words of the two documents are compared using WordNet which gives the degree of semantic similarity between the two documents. The degree of plagiarism is obtained after summarizing the above results which lies between 0 and 1.79. It has been observed that the semantic web technology has the potential of detecting plagiarism efficiently.

REFERENCES

[1] K.Sharma and B. Jindal, "An improved online plagiarism detection approach for semantic analysis using custom search engine," proc. IEEE 2016, International conference on computing for sustainable global development
[2] I. Smeureanu and B. Iancu, "Source code plagiarism detection method using protégé built ontologies," Informatica Economica vol. 17, issue 3, 2013.
[3] E. Al-Shamery and H. Gheni, "Plagiarism detection using semantic analysis," Indian journal of science and technology vol. 9(1), issue jan. 2016.
[4] J. Agarwal, R. Goudar, P. Kumar, N. Sharma, V. Parshav, R.Sharma, A. Shrivastava and S. Rao, "Intelligent plagiarism detection mechanism using semantic technology a different approach", international conference on advances in computing communications and informatics, IEEE, Los Alamitos, CA, 22-25 August,2013.
[5] J. Deepika, V. Archana, V. Bagyalakshmi, P. Preethi and G. Mahalakshmi,"A knowledge based approach to detection of idea plagiarism in online research publications", International Journal on internet and distributed computing system, vol.1 No.2, issue 2011.
[6] A. Osman, N.salim and M. Binwahlan, "Plagiarism detection using graph based representation", Journal of computing, vol. 2, issue 2, April 2010
[7] T. Chi, H. Wang, L. Liu, W. Song and C. Du, "Text similarity method based on ontology model", International conference on cloud computing and internet of things (CCIOT) 2014.
[8] T. Abdalla Elfadi Eisa and N. Salim, "Existing plagiarism detection techniques a systematic mapping of scholarly literature", Online Information Review Vol. 39 No. 3, 2015 pp. 383-400
[9] Wikipedia stop words list available at http://www.textfixer.com/tutorials/common-english-words.txt
[10] S. Manjula, K. Shet and U.Acharya, "Semantic plagiarism detection system using ontology mapping," Advanced computing international journal (ACIJ), vol.3, issue 3, may 2012.