Indian Journal of Science and Technology, *Vol 9(1), DOI: 10.17485/ijst/2016/v9i1/84235, January 2016*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Plagiarism Detection using Semantic Analysis

Eman Salih Al-Shamery[1] and Hadeel Qasem Gheni[2]*

[1]Information Technology College/ Babylon University, Iraq; emanalshamery@yahoo.com
[2]Software Department, Information Technology College/ Babylon University, Iraq; hadeelqasem84@Gmail.com

Abstract

The simplest description of a plagiarism is either a 'copy and paste' for a text even if the source was cited or a change in some words by taking the meaning without citing the source, where determining the meaning is the hardest and most complex task. Plagiarism can be seen as one of the cybercrime, similar to (computer viruses, computer hacking, spamming and the violation of copyrights), therefore, this subject has been interesting because it has become an important part of the ethics of scientific research. The increasing incidence of plagiarism in the higher education sector, which is considered acceptable behavior by some, since plagiarism saves time and effort, and gives better results, became a big problem faced by educational institutions. The main objective of this research is to find a suitable way to detect semantic plagiarism which occurs on the meaning and making use of synonyms and replace it instead of the original words. This research aims also to apply a pre-processing for the words of research by using tokenization and stop word removing processes, then tested whether the research enter under the specialization of computer science or not, where only such research will subject to semantic plagiarism detection by using WordNet. This research provides an effective way to detect semantic plagiarism for the written researches, especially by students who have a large plagiarism in their research.

Keywords: Plagiarism Detection, Semantic Plagiarism, Stop Words Removing, Tokenization, WordNet, WordNet Expansion

1. Introduction

Detect plagiarism has become a wide research area to reveal its types and so as to prevent the violation of rights, especially in education to prevent students from copyright infringement and to improve the educational level. Plagiarism is unacceptable use of the work of another author either as an accurate copy, or modify it a little bit[1]. Theft of the idea can be made fraudulently, especially if the source is not available to the public. The plagiarist steal the work of others, to be the owner and thus deprive the owner of the original work from this benefit. According to the online Dictionary of Merriam-Webster , the word "plagiarize" means to theft and pass off (ideas or words from another writing) as the owner, using (product of another) without citing to its source, clarify the idea by considering it new and innovative, while it is taken from present source[2].

In the era of communication, websites and e-books, plagiarism became very easy, which makes plagiarism very dangerous for the breadth of his chances, and severe unfaithfulness of intellectual property rights[3]. Plagiarism is a significant trouble[4]. The requirements of the academic work, especially research of it to write a thesis, its need to comparisons with previous research work to reveal the extent of literary plagiarism, so it is assumed that all universities need to measure the proportion of plagiarism and the scientific and literary thefts in the scientific researches to produce an original researches, as well as the student should not fear this type of program if he possessed the scientific secretariat, and documenting all sources, who takes them, this in order to avoid falling into the trap of scientific plagiarism.

Semantic plagiarism is a change in the meaning of words by taking synonyms of it, while retaining the positions of the words. There are a lot of theories in the field

**Author for correspondence*

of detection of plagiarism for the texts that contain significant changes in syntax and in meaning but mostly inadequate and inefficient, and this represents the biggest challenge in the detection of these changes, because it requires analysis of texts that carry similar meanings and making a decision whether there is a plagiarism or not[5]. Despite the fact that text similarities is a quick way to detect text plagiarism and has acceptable performance in situations that are copies of the original text as it is, can be easily deceived when working a simple paraphrasing. Because of this, the use of semantic relatedness will improve results by solving the mysterious and difficult issues of plagiarism[6]. For two texts, if we were able to extract the same semantic information, these two texts are considered semantically similar and can be interpreted as evidence that this is an issue of plagiarism.

Ordinary dictionaries cannot be suitable to be used to detect the complexities of meaning. Because the beneficial sentences consists of useful words, any system that process natural language should possess information about words and their meanings[7]. Similarity metrics determine the extent of the similarity of two concepts. There are various electronic dictionaries, lexical databases and thesauri today. WordNet is one of the largest and extreme wide used of these. It has been used in a variety of tasks such as the processing of natural language, which includes question answering and remove word meaning ambiguity. WordNet is a set of free software available that make it possible to measure the semantic similarity or correlation between a pair of concepts. It offers six metrics for similarity , and three metrics for relatedness, which is based on WordNet lexical database[8]. Synonymy is, of course, a lexical relation between word forms, in WordNet, the relationship that consider as the most important is the similarity that may be present in meanings. Two terms are considered synonymous when the replacement of each other does not change the meaning of the sentence in that place. Thus, according to this clarification, synonyms are scarce[9]. WordNet take into consideration the semantic areas of the word so that there is not only a text matching but looking for word meanings as well[10].

Many of the techniques proposed for detect semantic plagiarism in documents, [11]proposed a new method to detect paraphrased or translated text by a human by comparing the occurrences of citations in order to identify similarities. The most basic form is to measure the bibliographic coupling strength. [3]proposed a new method for semantic plagiarism using a synonym and antonym based framework to evaluate text similarity with respect to the similarity of content between the original and plagiarized document. [12]proposed a fuzzy system as a new method of plagiarism detection based on semantic based string similarity can handle external plagiarism detection as well as the fuzzy system can detect some means of obfuscation. [13]proposed a Semantic way for text clustering as a new method for plagiarism detection by using WordNet and lexical sequences to extract a group of related words semantically from texts that can represent the semantic content of the texts.

2. Implementation Methodology

In general, the proposed system consists of three main stages, each stage consists of many steps.

- Pre-processing Stage.
- Document Specialization Stage.
- Semantic Plagiarism Detection Stage.

Above all, building a database has been implemented to store documents according to special structure. This database contains two types of fields, fields to store an information about research and fields to store the research content. Fields for storing the information are: Public specialty of document (such as computer science, biology science.. etc.), specific specialization (such as networks, AI. etc.) and topic, while fields for storing the research content are: Title, keywords, abstract and finally the rest of the text. Whenever the database is of a large size, the more increased accuracy in matching data. Figure 1 explain the structure of database.

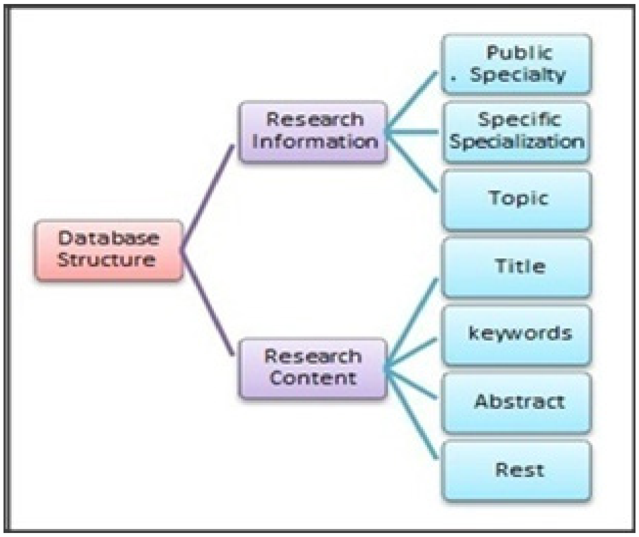

Figure 1. Database Structure.

2.1 Pre-processing Stage

This phase consists of several steps as shown in Figure 2.

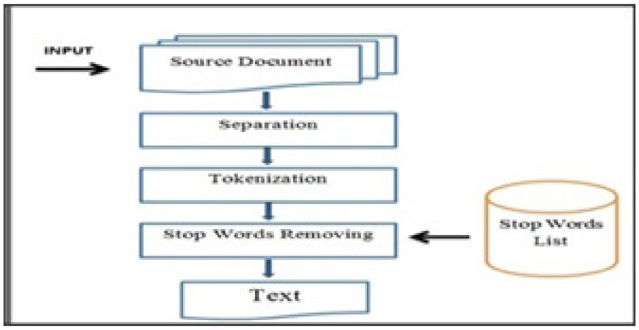

Figure 2. Pre-processing Stage.

2.1.1 Separation

In this step, the text of input document is isolated from the references mentioned therein. Separating the references from the text can be manually or programmatically.

2.1.2 Tokenization

The text of the document that consists of paragraphs is divided into set of tokens in a process called tokenization. The output of this stage is to convert the file content to an individual words. After that, a deletion process will occur to the delimiters which may be a companion to these words. The tokenization and delimiter deletion algorithm can be described in 5 steps.

```
Step1: Declare String array text[],text2[], Declare String line.
Initialize Integer C1_text, C2_text2, sum_text,sum2_text to 0
Step2: Set line = in.readLine();            // to fetch line from file
Step3: Do
WHILE line is not equal to null
set text[]=line.split(" "); // split the line based on space increment sum_text; // sum = length of array text[].
ENDWHILE; Set next line by : line = in.readLine(); // fetch the next line
Until (end of file). // Now, all the lines are in array text[].
Step4: WHILE C1_text  < sum_text
set text2[C2_text2]= text[C1_text].replaceAll("[\\W]", ""); // delete the delimiters
Increment C1_text
ENDWHILE;
C2_text2=sum2_text2; // sum2= length of array text2
Step5: Print text2[C2_text2] // text2[]= individual words without delimiters.
```

The delimiters that deleted from the text are explained in table_1.

Table 1. Delimiters

{	}	[	]
\	|	“	‘
:	;	+	=
_	-	)	(
*	&	^	%
$	#	@	!
~	?	/	>
<	.	,	؟

2.1.3 Stop Words Removing

Stop words are words that repeated frequently in the eEnglish language, but do not carry any information. These words may be kind of pronouns, conjunctions and prepositions. The output of this stage is a text free of stop word, finally puts all letters in lower case. Table_2 shown stop words that removed from the text.

The output of pre-processing stage is a text ready to check against semantic plagiarism.

2.2 Document Disciplinary

Before detect the semantic plagiarism, a process of identifying the specialist of document is done to detect plagiarism only for documents that fall within the specialty of computer science, while the documents with other disciplines will not subject to plagiarism detection. Figure 3 illustrates this process.

2.2.1 Word Frequency

After pre-processing stage, The occurrence of each word in the input document will computed according to how many times it appears in document.

2.2.2 Descending Order

Frequencies that found in the previous step will arranged in descending order.

2.2.3 N Specification

At this stage, N was determined within the program that representing the highest frequencies will be taken from the total number of it.

Table 2. Stop Words Removing

A	am	an	etc.	e.g.	If	for	from
in	on	Is	are	was	were	has	have
be	he	she	him	her	It	that	this
those	do	does	did	didn't	doesn't	didn't	they
them	these	I	me	my	mine	you	your
yours	its	we	us	our	ours	their	theirs
hers	and	or	other	of	off	aren't	but
by	can	cannot	can't	could	couldn't	shall	should
shouldn't	to	too	up	very	we'd	we'll	we're
we've	what	what's	when	when's	where	where's	which
while	who	who's	whom	why	why's	with	won't
would	wouldn't	whole	over	then	than	therefore	yourself
yourselves	you've	you're	once	had	hadn't	hasn't	hasn't
having	I'd	I'll	I'm	no	into	again	against
all	anyone	any	above	about	before	after	below
more	most	mustn't	let's	nor	thus	the	much
many	like	likely	few	little	so	same	de
Some	Something	Nothing	anything	usually	always	at	as
i.e.	Inc.	Ltd	Re	miles	km	http://	per
A	B	C	D	E	F	G	H
I	J	K	L	M	N	O	P
Q	R	S	T	U	V	W	Q
Y	Z	a	b	c	d	E	f
g	h	i	J	k	l	m	n
O	p	q	r	s	T	u	v
w	x	Y	z	1	2	3	4
5	6	7	8	9	10	11	12
13	14	15	16	17	18	19	20
1-	2-	3-	4-	5-	6-	7-	8-
9-	10-	11-	12-	13-	14-	15-	16-

2.2.4 Word (N)

Words that represent the highest (N) frequencies will take a side to reveal the correlation of source document with the computer science fields.

2.2.5 Specialism of Documents

The aim of this step is to accept the documents that associated only with computer sciences, this mean, if the source document associated with other sciences, will be rejected and cannot complete the work with it. Words that have already been taken will be matched with a new database contain a table including all the fields that are related to computer science, if one of the words resulted from previous step match one of the database fields then the document is relevant to computer science and will subject to semantic plagiarism detection, while if the words doesn't match any of computer science fields then the document will be unrelated to computer science and will reject and stop working. Specialist of document algorithm can be described in 7 steps.

Step1: Build a database containing one table involving the fields of computer science.

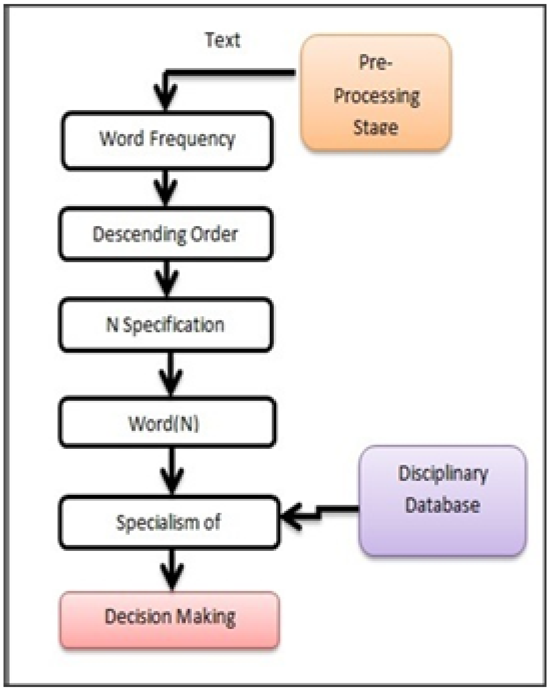

Figure 3. Identifying the Document Disciplinary Process.

Step2: Extract the plain text from the input research.

Step3: Removes all characters except letters and puts everything in lowercase.

Step4: Compute the frequency for each word in the research.

Step5: Filtering the frequencies by taking the top 30 one.

Step6: Taking the words that relates to these 30 frequency.

```
1) For I= first field to the last field in a database table.
Matching the 30 word.
2) If the result = 0 then
No matching, reject the document.
else
Document  is relevant with at least one of the fields.
End if
```

Step7: Display the fields that are relevant to the research.

2.2.6 *Decision Making*

Finally, the fields that are related to this document will be displayed and continue working.

2.3 Semantic Plagiarism Detection

Then, to help detecting semantic plagiarism, we propose to use semantic similarity between documents based on information extracting techniques. semantic plagiarism will be detect based on WordNet, it has several steps as shown in Figure 4.

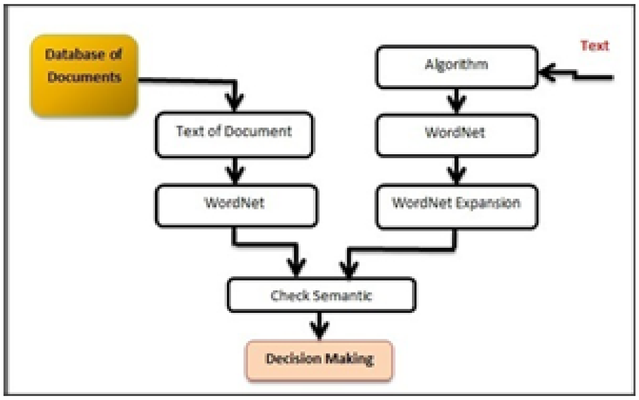

Figure 4. Semantic Plagiarism Detection.

2.3.1 *Text*

If the document passed from the threshold of specialty test of computer science, the text will be taken once again to complement this work.

2.3.2 *Algorithm*

Only the algorithm or the proposed system contained in the document is subject to detect semantic plagiarism.

2.3.3 *WordNet*

After taking the text that was specified in the previous step, to determine the extent of the semantic plagiarism, synonyms for each word is to find using WordNet. Every word in the specified text will be extracted its synonyms. These synonyms will be considered as appearance of the word itself when used to detect plagiarism.

2.3.4 *WordNet Expansion*

At this stage, WordNet expansion has been proposed by specific words doesn't exist in its dictionary. words were clarified in table_3.

2.3.5 *Documents of Database*

At this stage, the documents stored in the database will be withdrawn one after another, for these documents, the text is taken entirely not just a specific text in it, to the possibility of plagiarize a text exist in different places of database document and put it in another place of source document, these places may be abstract, results,

Table 3. WordNet Expansion

WORD	SYNONYMS
Method	Algorithm, Tool, Model, System, Steps, Approach, Paradigm, Scheme, Technique.
Architecture	Block diagram, Flowchart, Framework, Structure
Proposed	Introduced, Employed, Exploited, Suggested, Reviewed, Developed, Applied .
Develop	New, Novel , Propose, Suggest
High	Promised , Excellent .

conclusions, or the proposed system. After that, WordNet will be applied on these documents to find the synonyms.

2.3.6 Check Semantic Plagiarism

Most important step in our work is plagiarism detection process that has been implemented based on the meaning of words and their positions, if the plagiarist change the meaning of words but the words remain in the same sites where in the original text, this is a semantic plagiarism and this is what has been discussed. In this work, the word in the algorithm of the source document will be searched for in the document drawn from the database. if this word or one of its synonyms is found then will checked the words after the first word in the same context and if it in the same locations in the document drawn, this will be considered as semantic plagiarism, but if the words or its synonyms are found, but in a different order of what exists in the document drawn then this is not considered as semantic plagiarism.

3. Discussion

The proposed system steps in this research was applied to many of the documents using the Java NetBeans IDE 8.0.2 language. No matter what the type of document, for any case, it will be converted to Text to be handled according to the system. Meanwhile, the database was built by MySQL workbench 6.3 CE program and the connector between java program and database is MySQL Community Server (GPL) version of 5.7.9 in 3306 port.

The database has a capacity of storage equal to 500,000 documents. Size of each document can range between 1-50 sheet. The database is made up of several fields in which to store documents, firstly, an general information about the document to be stored was entered, this information is public specialization, then specific specialization and finally topic. The type of fields that contain the text of this information is Varchar (255 char). Secondly, the contents of the document are stored in fields, which are, title, abstract, keywords and the rest of the content. The type of title and keywords fields is Varchar, while the type of abstract field is TEXT(65,535 char) and finally the type of rest field that contain the remaining text of the file is LONG TEXT (4,294,967,295 char).

In this research, work was based on WordNet to give the synonyms to the words in the document, the synonyms will consider as occurrence for the same word. The input is document file and the output is a report about the semantic plagiarism that occurs in the algorithm or proposed system between the source document and other documents stored in the database. Table_4 shown the semantic plagiarism percentages for one source document with 100 database documents.

Accordingly to the values mentioned in Table_4, it was found a semantic plagiarism in algorithms or the proposed system of document when compared with the documents of database. Detect semantic plagiarism is by finding synonyms for words and inspect their sites, if there was stability in sites and the plagiarist has only replaced the word in one of its synonyms, this is considered as a semantic plagiarism. The run time required to get the matching result is 6 seconds. The following diagram shows the percentages for semantic plagiarism detection process.

According to the diagram shown in Figure 5, semantic plagiarism has been detected in a sample of 100 documents stored in the database with different percentages, where the highest percentage is 0.53% and least percentage is 0%.

4. Conclusion

This paper, describes an approach to detect semantic plagiarism which occurs in researches by using WordNet. In this approach, WordNet has proven as an effective way to identify the semantic plagiarism by given the synonyms of words in the document then detect the plagiarism, Then, is to know the change in the words locations that have been changed by other synonymous words. If there is no change in words locations as it exists in the database

Table 4. Percentage of Semantic Plagiarism

Doc No.	Plagiarism. Percentage	Doc No.	Plagiarism. Percentage	Doc No.	Plagiarism. Percentage	Doc No.	Plagiarism. Percentage
1	0.07	26	0.05	51	0.03	76	0.04
2	0.14	27	0.03	52	0.2	77	0.08
3	0.1	28	0.2	53	0.1	78	0
4	0.05	29	0.02	54	0.04	79	0
5	0.2	30	0	55	0.53	80	0
6	0.4	31	0	56	0.3	81	0
7	0.09	32	0	57	0.04	82	0
8	0.02	33	0	58	0.06	83	0.05
9	0.19	34	0	59	0.05	84	0.3
10	0.12	35	0.2	60	0.09	85	0.4
11	0	36	0.3	61	0	86	0.3
12	0	37	0.4	62	0	87	0.49
13	0	38	0.1	63	0	88	0.2
14	0	39	0.03	64	0.3	89	0
15	0	40	0.07	65	0.2	90	0.5
16	0	41	0.04	66	0.3	91	0.07
17	0.01	42	0.08	67	0.1	92	0.05
18	0.09	43	0	68	0.04	93	0.08
19	0.03	44	0.09	69	0.06	94	0
20	0	45	0	70	0.06	95	0.01
21	0	46	0	71	0.07	96	0.03
22	0	47	0	72	0.03	97	0.07
23	0	48	0.03	73	0.01	98	0.2
24	0	49	0.02	74	0.09	99	0.1
25	0.1	50	0	75	0.01	100	0.01

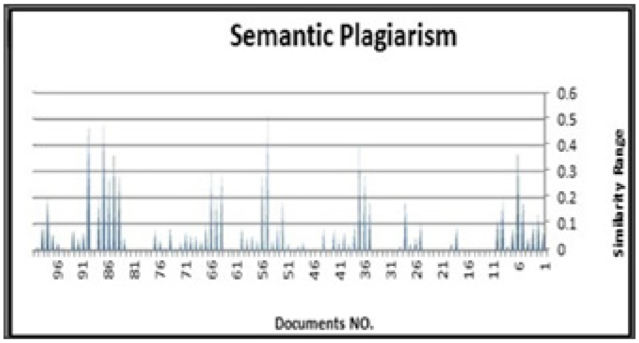

Figure 5. Percentages of Semantic Plagiarism Detection.

document, then, there is a semantic plagiarism. In this work, the WordNet has expanded to be an entrance to other research where it has been added a several meanings of certain words and included in WordNet to be utilized to detect the semantic plagiarism.

5. References

1. Stamatatos E. Intrinsic plagiarism detection using character n-gram profiles. In: Stein, Rosso, Stamatatos, Koppel, Agirre, editors. PAN'09. 2009; 38–46.
2. Ison DC. Does the online environment promote plagiarism? A comparative study of dissertations from brick-and-mortar versus online institutions. MERLOT Journal of Online Learning and Teaching. 2014; 10(2):272–82.
3. Shams K. Plagiarism detection using semantic analysis. Diss. BRAC University: Dhaka, Bangladesh; 2010.
4. Niezgoda S,Way TP. SNITCH: a software tool for detecting cut and paste plagiarism. ACM SIGCSE Bulletin. 2006; 38(1):51–5.

5. Chong MYM. A study on plagiarism detection and plagiarism direction identification using natural language processing techniques. University of Wolverhampton: England. 2013.
6. Tsatsaronis G, Varlamis I, Giannakoulopoulos A, Kanellopoulos N. Identifying free text plagiarism based on semantic similarity. Proceedings of the 4th International Plagiarism Conference. 2010.
7. Millar GA. WordNet: a lexical database for English. Communications of the ACM.1995; 38(11):39–41.
8. Pedersen T, Siddharth P, Jason M. WordNet: Similarity: measuring the relatedness of concepts. Proceeding Demonstration papers at HLT-NAACL Demonstration'04. 2004; 38–41.
9. Miller GA, Beckwith R, Fellbaum C, Gross D, Miller K. Introduction to wordnet: An on-line lexical database. International journal of lexicography, 1990; 3(4): 235–44.
10. Chen C-Y, Yeh J-Y, Ke H-R. Plagiarism Detection using ROUGE and WordNet. Journal of Computing. 2010; 2(3):34–44.
11. Gipp B. Citation-based Plagiarism Detection–Idea, Implementation and Evaluation. 2011; 1–11.
12. Alzahrani S, Salim N. Fuzzy semantic-based string similarity for extrinsic plagiarism detection. 2010; 1–7.
13. Wei T, Lu Y, Chang H, Zhou Q, Bao X. A semantic approach for text clustering using WordNet and lexical chains. Expert Systems with Applications, 2015; 42(4):2264–75.