Asian Journal of Information Technology 15 (16): 2743-2746, 2016
ISSN: 1682-3915

Paraphrasing: Solutionto The Problem of Plagiarism

Ahmed Mohammed Hussein
Department of Computer Science, College of Science for Women,
University of Babylon, Babylon, Iraq

Abstract: Academic production mostly relies on external sources have been published used by the researcher in order to support his ideas, opinions and the results that may be reached by. And so the use of external sources is acceptable automatically but what is unacceptable is the failure of the researcher mentioning those sources which leads to plagiarism. The plagiarism is stealing the ideas of others and a violation where the university take it seriously in all cases. The best way to avoid plagiarism is rewriting texts that quoted from another source in a way that allows writing it of a new form as well asthe reference to the sources this so-called paraphrasing. In this study, we have a way to detect the paraphrasing used on the texts as well as determine the percentage to disclose the amount of the change on the texts. Results proved that the researcher is to paraphrase the texts taken from other sources at different rates of which exceeded 45% including good paraphrasing by typing text in a new way in addition to changing the words sites but some are very simple change in meaning and location.

Key words: Plagiarism, paraphrasing detection, WordNet, exceeded, rewriting texts

INTRODUCTION

As the industrial revolution changed the way people live and the way they work, the computers development and the expansion of the internet and search engines have changed the way of thinking.Thus, the enormous amounts of information are now accessible to all people through access to the Internet with no need to magnify skills for that. As a result, the traditional ways to search for information through the books have become a thing of the past (Sraka and Kaucic, 2009).

Internet has given the possibility for students to find a lot of examples of programs and source code so these resources has become an essential resource for plagiarism (Ohno *et al*., 2011).

Plagiarism is taking a duplicate copy or slightly modified version of the work of another researcher without permission to do so (Stamatatos, 2009).

Changing the words of the sentences belonging to the original text by reformulating or modifying considered as plagiarism when the writer keeps on the locations of these words as they are while when the writer rewrite the words or taking the synonym of them. Change their positions and certainly cite to the original source, this is what is considered acceptable which is called paraphrasing.

Paraphrasing is rewriting ideas or rewrite the words using other words special for the writer. Synonyms of words can be use or using another words as well as changing the structure of phrases where it can be change the order of words that existing in the original text to create a personal structure of writer.

Although, the text similarities is the fastest way to detect textual plagiarism and which is successful in cases where the text is an exact copy of the original text, it can be easily duped when make a simple paraphrasing for some words (Shamery *et al*., 2016).

Another term is quoting, take the text from another source and put it in the study of the writer by putting quotation marks, this is what will distinguish the text of the writer from text belonging to another source. In addition to quotation marks, it must cite the source of that text. Paraphrasing is the best than quoting because it helps the writer to understand the full meaning of the text he want to quote from it as well as not to exceeded to copy many of texts.

The successful paraphrasing require using as few words of the original text as possible where the writer must read the original text repeatedly until he understanding it well then write the text in his own words and do not forget to cite the original source . Without a successful paraphrasing, the text can be interpret as plagiarism.

In this research, WordNet used to give synonyms for words. WordNet is a database for English-language and can be consider as comprehensive database of dictionary and thesaurus, containing metrics for similarity and relatedness and the most importance metric between the forms of words is synonyms. When to replace a term with another term and does not alter the meaning of the sentence in that location, the two terms are considered synonymous (Shamery *et al*., 2016).

Related works: Recently this subject is of growing interest where detection of paraphrasing used in many applications. Many of the techniques proposed for detect paraphrasing in documents (Brun *et al*., 2003) built a system for document processing and the output is a normalized representation of some chosen knowledge, the analysis phase can be view as a paraphrase detection stage.

Salvador *et al*., 2014) use an approach based on knowledge graph for getting and comparing models of documents in various languages to enhance the detection of paraphrasing.

Berant use a way to generate an inevitable set of logical candidate forms with the realization canon in natural language for each. Then, use a paraphrase model to choose the realization that best paraphrases the inputs and it produces a logical format corresponding.

Boonthum *et al*. (2003) described the target and the need to recognize paraphrasing strategy as well as it focused on the definition of paraphrasing and patterns of discrimination, also discussed multiple representations of knowledge could be used to recognize the paraphrasing.

Olivares *et al*. (2013) analysis feature is perform to accomplish paraphrase recognition and recognition of textual entailment experimenting with a mixture of various natural language processing mechanism.

Socher introduced technology to detect the paraphrasing relying on Recursive Auto Encoders (RAE) and this depends on a new unfolding objective and vectors of learn feature for the phrases in the syntactic trees. These features used to measure the similarity of word and phrase between two sentences.

MATERIALS AND METHODS

Implementation methodology: The proposed system is working to identify the texts that the researcher taken from other sources and then make a paraphrasing for it.For the purpose of comparison the input file with several files to find and detect paraphrasing of texts written in it, a database built for storing research and then retrieved for the purpose of comparison. Figure 1 illustrates the proposed system).

Pre-processing stage: When submitting the input file into the system, the first operation will take place is the pre-processing phase which will include separate the paragraphs to an individual words and then delete symbols, digits and everything except the letters which are called delimiters.

Wordnet application: The second operation is WordNet application that mean extract the synonyms for each word written in the original research and be stored temporarily for comparing it with the words of research stored in the database. After that, the research stored in the database are pulled one after the other for the purpose of comparing it with the input research . Each research will be going through the same processes, pre-processing and the application of WordNet.

Paraphrasing detection: The third operation that is the basic in our research is paraphrasing detection, in this process, the word, its synonyms and its site will take into consideration, paraphrasing of the text happened when changing a word by its synonyms and moved from its site. Writer can performed a slightly changing in the sites or make full paraphrasing where he is reading and understanding the text taken from another source repeatedly then rewritten in own style.

Similarity ratios calculations: After that in the fourth operation, words that have been discovered that it changed by one of its synonyms and change its site will be calculated and divided by the total number of words to find the proportion of paraphrasing in the input research.

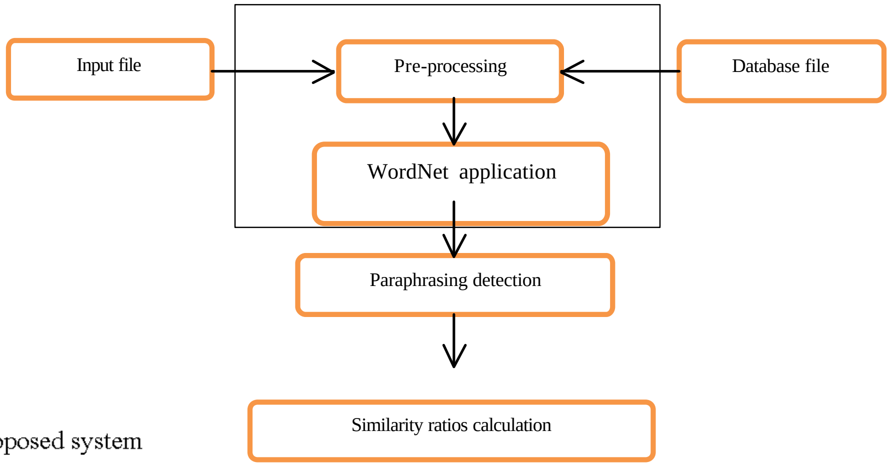

Fig. 1: The proposed system

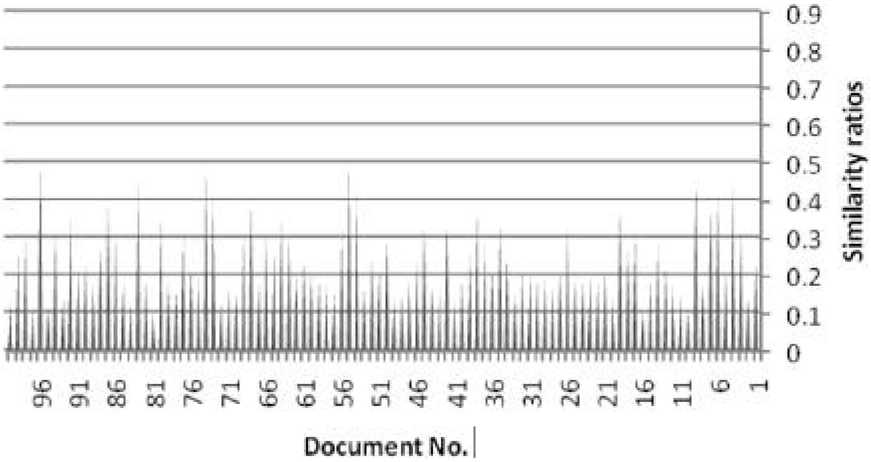

Fig. 2 : Paraphrasing detection process

Paragraphs that have been paraphrase will be display as well as display the quoted text.

The following is an algorithm showing the steps of paraphrasing detection and determination of the percentage of paraphrasing.

```
Name : Paraphrasing detection
Input : Source document
Output : Paraphrase texts and percentage
begin
Step1:For each of the source document and database document:
Extract the plain text from the source document.
Implement the pre-processing process.
Get synonyms for each word by applying WordNet.
Step2:For each word in the source document, find it or one of its synonyms in the text of database document.
If found:
Compared the two sentences (source document and database sentences) in terms of synonyms and the order of locations.
Calculate the number of similar words.
Calculate the total number of words in a sentence.
Find the ratio by dividingthe number of similarities on the total number of words.
Step3:Display the paraphrasing texts and similarity percentage.
End
```

RESULTS AND DISCUSSION

In our research, WordNet programused to take advantage of synonyms for words.The input research is compared with a set of research buffered in a database, where it is compared the words of paragraphs of the input research with the words of the paragraphs of the research in the database and taking into consideration the synonyms for these words.

The program built using Java NetBeans IDE 8.0.2language and the database built by MySQL workbench 6.3 CE program with MySQL Community Server (GPL) version of 5.7.9 connector in 3307 port.Table 1 shows the results obtained from the test set which included a total of 100 documents. It has been applied detect of paraphrasing to a one document and compared with the rest of documents.

Table 1: Paraphrasing percentage for 100 document

	Paraphrasing		Paraphrasing		Paraphrasing
Doc No.	(%)	Doc No.	(%)	Doc No.	(%)
1	0.23	35	0.32	69	0.32
2	0.14	36	0.22	70	0.15
3	0.32	37	0.28	71	0.17
4	0.45	38	0.38	72	0.13
5	0.21	39	0.26	73	0.41
6	0.42	40	0.90	74	0.49
7	0.38	41	0.13	75	0.19
8	0.17	42	0.34	76	0.22
9	0.45	43	0.15	77	0.31
10	0.11	44	0.90	78	0.16
11	0.15	45	0.33	79	0.18
12	0.17	46	0.24	80	0.37
13	0.23	47	0.19	81	0.10
14	0.28	48	0.15	82	0.19
15	0.19	49	0.14	83	0.44
16	0.10	50	0.29	84	0.13
17	0.31	51	0.22	85	0.18
18	0.27	52	0.24	86	0.31
19	0.38	53	0.17	87	0.39
20	0.14	54	0.42	88	0.27
21	0.22	55	0.51	89	0.18
22	0.18	56	0.32	90	0.23
23	0.20	57	0.16	91	0.22
24	0.8	58	0.18	92	0.36
25	0.19	59	0.19	93	0.14
26	0.33	60	0.20	94	0.31
27	0.21	61	0.24	95	0.12
28	0.17	62	0.19	96	0.48
29	0.19	63	0.31	97	0.11
30	0.9	64	0.35	98	0.29
31	0.21	65	0.27	99	0.27
32	0.22	66	0.31	100	0.13
33	0.8	67	0.18		
34	0.25	68	0.39		

Results shown in the table above shows the presence of varying proportions of paraphrasing, including small percentages and ratios which almost half the size of the document. The diagram bellow shows the percentages for paraphrasing detection process (Fig. 2)

CONCLUSION

In this study, we have offered a method that can be usedto reveal the paraphrasing in the texts with the help of WordNet.In this research, has been relying on the WordNet, where we could get the synonyms of the word through it.Results proved that the researchers have done a paraphrasing for the text quoted from another source but in different proportions where they re-writes the words in a different way depending on the synonyms as well as change the sites of the words in order to avoid plagiarism and ensure the rights of researchers.

REFERENCES

Boonthum, C., S. Toida and I. Levinstein, 2003. Paraphrasing recognition through conceptual graphs. Computer Science Department, Old Dominion University, Norfolk, Virginia.

Ohno, A. and H. Murao, 2011. A two-step in-class source code plagiarism detection method utilizing improved CM algorithm and SIM. Int. J. Innovative Comput. Inf. Control, 7: 4729-4739.

Olivares, A.S., A. Garcia and H. Calvo, 2013. Feature analysis for paraphrase recognition and textual entailment. Res. Comput. Sci., 70: 119-144.

Salvador, M.F., P. Gupta and P. Rosso, 2014. Knowledge graphs as context models: Improving the detection of cross-language plagiarism with paraphrasing. In: Bridging Between Information Retrieval and Databases, Ferro, N., (Ed). Springer, Berlin, Germany, ISBN:978-3-642-54798-0, pp: 227-236.

Shamery, A., S. Eman and Q.G. Hadeel, 2016. Plagiarism detection using semantic analysis. Indian J. Sci. Technol. Vol.9,

Sraka, D. and B. Kaucic, 2009. Source code plagiarism. Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces, June 22-25, 2009, IEEE, New York, USA., ISBN:978-953-7138-15-8, pp: 461-466.

Stamatatos, E., 2009. Intrinsic plagiarism detection using character n-gram profiles. Threshold, 2: 1-500.