ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 11, November 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Survey on Different Obfuscated Plagiarism Detection Methods

Vijay S Mhaskar[1]
Department of Computer Engineering
Shree Ramchandra College of Engineering, Lonikand
Pune – India

Prof. Baban Thombre[2]
Department of Computer Engineering
Shree Ramchandra College of Engineering, Lonikand
Pune – India

Abstract: ***Plagiarism is using other people's work or ideas and using them off as one's own without acknowledging them back. Identifying duplicated and plagiarized passages of text is becoming popular area of research. Issue of plagiarism is becoming worse with easily available resources on web. Nature of plagiarism ranges from copying texts to adopting ideas, without giving credit to its originator, changing texts into semantically equivalent but with different words and organization, shortening texts to summarize concept, and adopting ideas and important contributions of others. Several studies also suggest that plagiarism is very common among medical students. This paper presents a survey on different plagiarism detection techniques; feature provided by them and highlights current limitations of those systems.***

Keywords: Information Retrieval; MEDLINE; Plagiarism Detection; Query Expansion; Search Engine.

I. INTRODUCTION

Plagiarism is the act of using another person's words or ideas without giving credit to that person. Identifying duplicated and plagiarized passages of text is becoming popular area of research. Plagiarism is an important issue in every academic and research institutes and this situation is becoming worse with easily available online resources. MEDLINE contains a more than 22 million publications in the area of medicine and related fields. New publications are getting added continuously which makes difficult for individuals or groups to keep eye on the information contained within it. Most of the authors know that plagiarism is unethical publication practice. Yet, it is a serious problem in the medical writing arena. There are different types of plagiarisms which are plagiarism of text, plagiarism of ideas, mosaic plagiarism, self-plagiarism, and duplicate publication [4].

II. TYPES OF PLAGIARISM

Alzahrani, Salha M., Naomie Salim, and Ajith Abraham [3] have conducted a qualitative study at the University of Technology Malaysia to understand academician's experience, who faces plagiarism, to pursue in-depth information around the offence, and to get the in depth knowledge about current plagiarism practices committed by the students. This data was collected by conducting interviews of faculty members with 10-20 years of teaching expertise. Questions were mainly based on different plagiarism practices by the students. The output of this study is a new taxonomy of plagiarism that comprehensively relates different types, as shown in Fig. 1. The taxonomy divides plagiarism into two typical types based on the plagiarist's behaviour.

1. **Literal plagiarism** - Literal plagiarism is a common and major practice wherein plagiarists do not spend much time in hiding the academic crime they committed. For example, they simply copy and paste the text from the Internet. Aside from few alterations in the original text.
2. **Intelligent plagiarism** - Intelligent Plagiarism is a serious academic dishonesty wherein plagiarists try to deceive readers by changing the contributions of others to appear as their own. Intelligent plagiarists try to hide, obfuscate, and change the original work in various intelligent ways, including text manipulation, translation, and idea adoption.

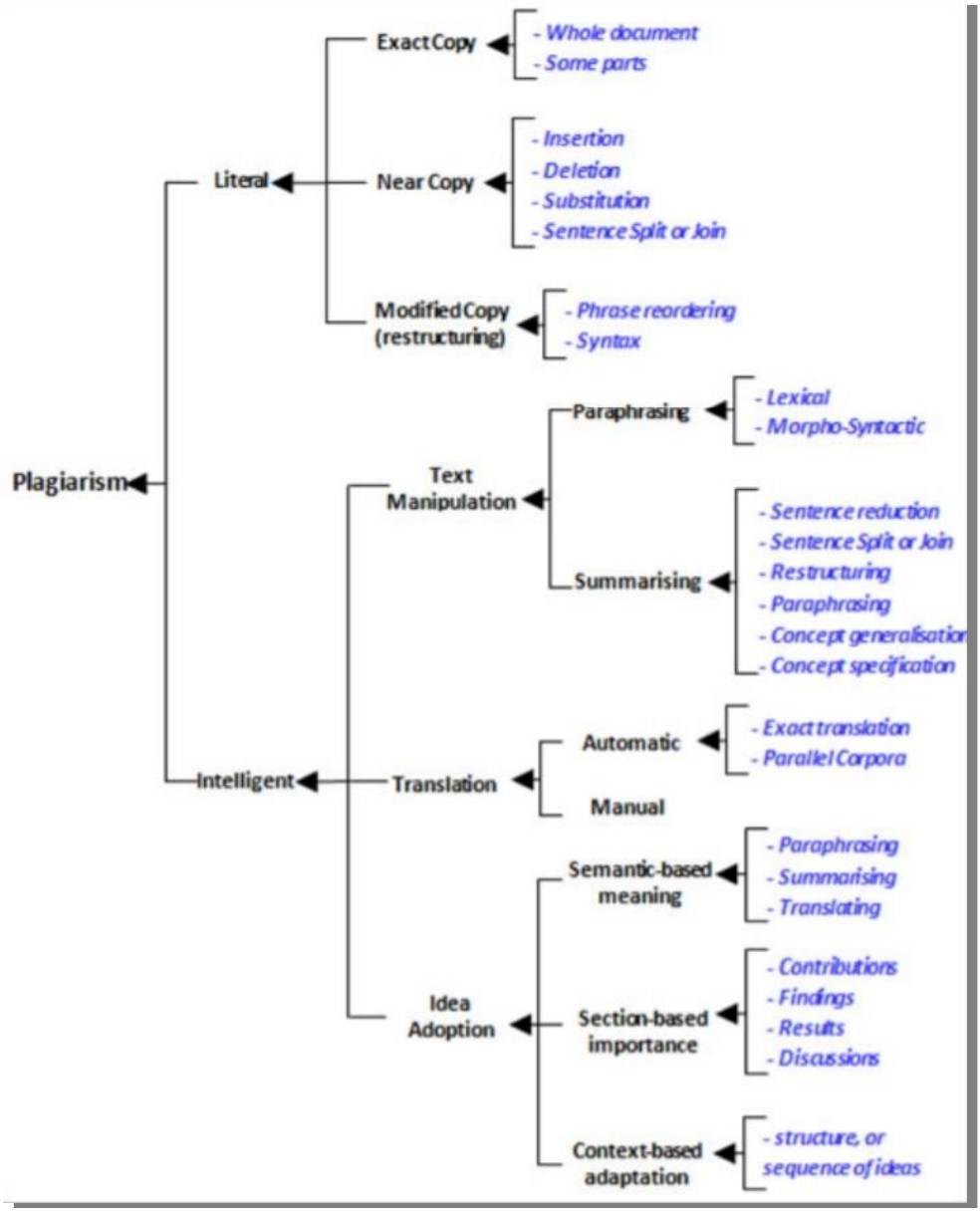

Fig. 1 Taxonomy of Plagiarism

III. LITERATURE SURVEY

In plagiarism detection systems, mainly two important problems are considered: the problem of candidate documents selection that are globally similar to a document which is under investigation, and the problem of comparing of document under investigation and its candidates to pinpoint plagiarized fragments in detail.

In [1] Rao Muhammad Adeel Nawab, Mark Stevenson and Paul Clough proposed an approach to find plagiarism in MEDLINE using query expansion techniques. Query expansion is performed using the ULMS Metathesaurus to deal with situations in which original documents are obfuscated. It mainly focuses on cases where plagiarised text has been highly obfuscated which presents major challenge to automated plagiarism detection systems. Evaluation was carried out using the MEDLINE Corpus, which contains potential real cases of plagiarism. Results show that the IR-based approach using query expansion outperforms a state-of-the-art approach. The IR-based approach proposed here achieves higher results than the Kullback-Leibler Distance approach. Although it is expected that performance will drop when the entire MEDLINE database is used.

In [2] Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, Paolo Rosso described approaches for producing exact and modified copies. Detecting plagiarism which involves little or no modifications of the original document is straightforward.

However real examples where source text was rewritten using Anti Anti Plagiarism systems. This paraphrased passage was analysed by two well-known commercial plagiarism detection services and both are failed to detect plagiarism.

In [3], Alzahrani, Salha M., Naomie Salim, and Ajith Abraham proposed detail taxonomy, approaches used and semantic framework of the tool. After detailed Analysis of various methods, authors suggest that semantic and Fuzzy Based method can provide better results. Both semantic and Fuzzy are challenging areas. Since no standard Fuzzy data-set is available [5] to find fuzzy words whereas in case of semantic detection it is difficult to represent semantic of sentence. They also explained various other approaches which are used to detect plagiarism, It covers simple lexical methods to complex semantic methods. He also concluded that current plagiarism tools for educational institutions, academicians, and publishers limited to word to word plagiarism and only some instances of it. They do not cover adapting ideas from others. Now plagiarism has become more sophisticated, idea plagiarism is a key academic problem and should be addressed in future research. He also proposed structural features and contextual information with efficient STRUC-based methods to detect section based importance and context based adaption idea plagiarism.

In [6]. Chong, L.Specia, and R. Mitkov Proposed a framework for external plagiarism detection in which a number of NLP techniques which includes Tokenization, sentence segmentation, Part Of Speech (POS) tagging, chunking and dependency parsing etc are applied to process a set of suspicious and original documents, not only to analyse strings but also the structure of the text, using resources to account for text relations. Initial results obtained with a corpus of plagiarised short paragraphs have showed that NLP techniques improve the accuracy of existing approaches. This paper provides an insight of how NLP techniques are capable of improving current plagiarism detection methodologies, and also invokes further investigation on applying high level NLP approaches to develop a better methodology.

In [8] Alberto Barrón-Cedeño, Paolo Rosso, José-Miguel Benedí proposed an approach which retrieves candidate documents using the Kullback-Leibler Symmetric Distance method. Documents are modelled as probability distributions and compared using KLδ. Documents are converted into probability distributions by removing stop words, stemming and then computing tf.idf weights for the remaining word unigrams. Results showed that the overall accuracy and speed of the plagiarism detection system improved by applying the Kullback-Leibler Symmetric Distance to reduce the plagiarism detection search space.

In [9] J. Lewis, S. Ossowski, J. Hicks, M. Errami, and H. Garner proposed a vector-based text similarity search algorithm to identify highly similar citation pairs in MEDLINE. They have also demonstrated that proposed novel text similarity algorithm, when coupled with word-vector approaches, is a highly effective alternative to traditional techniques, enabling us to offer the biomedical research community a literature search tool which is optimized, simple and free. A limitation of this is that it is unable to identify similar MEDLINE citations when the original text has been substantially altered, such as by paraphrasing or replacing words with synonyms [10].

IV. Research Direction

Existing systems performs poorly when it comes to multiple complex matching operations on large document collections such as MEDLINE. We need to address platform scalability problem in plagiarism detection system. The system that we are using for to detect suspicious documents should be scalable enough to handle large document collection as well as complex query operations. Information retrieval system based on popular open source search platform Apache Solr which is proven to be scalable for billions of documents. Apache Solr is based on popular Apache-Lucene search library. Lucene has its own relevancy framework which computes document score on various factors such as tf, idf, coord, length norm etc. Also while searching for suspicious documents in MEDLINE corpus if only exact phrase match technique is used then It may fail to identify the document similarity if original text has been rewritten which most of the plagiarist attempt. This needs to handled using multiple match types along with phrase match. Finally, while generating final ranked list of documents existing system

uses combSUM as a rank fusion technique which will not have influence of match type of search query. We need to merge result set from each query type by considering weight of each query type.

V. CONCLUSION

As digital document is being easily available on internet many times it happens that people makes the near copies of the original document without giving credit to original author. Due to this reason use of plagiarism detection systems has become very important practice in any research. In this paper a survey on plagiarism detection systems has been introduced. We have studied various approaches that are available to the problem of candidate document selection and query expansion for extrinsic plagiarism detection. We have also presented list of advantages and disadvantages of the latest and the important effective methods used or developed in automatic plagiarism detection, according to their result.

ACKNOWLEDGEMENT

I am profoundly grateful to Prof. Baban Thombre for his guidance and continuous encouragement throughout to see that this paper rights its target since its commencement to its completion. I would like to express deepest appreciation towards his invaluable guidance and support me in completing this paper. Also I must express my sincere gratitude to all the staff members of Computer Department, SRCOE Lonikand, Pune who helped me directly or indirectly during this course of work.

References

1. R. Muhammad Adeel Nawab; M. Stevenson; P. Clough, “An IR-based Approach Utilising Query Expansion for Plagiarism Detection in MEDLINE," in IEEE/ACM transactions on Computational Biology and Bioinformatics , vol.PP, no.99, pp.1-1, doi: 10.1109/TCBB.2016.2542803
2. M. Potthast, A. Barr´on-Cede˜no, B. Stein, and P. Rosso, “Crosslanguage plagiarism detection,” Lang. Resour. Eval., vol. 45, no. 1, pp. 45–62, Mar. 2011. [Online]. Available: http://dx.doi.org/10.1007/ s10579-009-9114-z
3. Alzahrani, Salha M., Naomie Salim, and Ajith Abraham. “Understanding plagiarism linguistic patterns, textual features, and detection methods.” Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 42.2 (2012): 133- 149.
4. Das N, Panjabi M. Plagiarism: Why is it such a big issue for medical writers? Perspectives in Clinical Research. 2011;2(2):67-71. doi:10.4103/2229-3485.80370.
5. Chandran, David, Keeley Crockett, David Mclean, and Zuhair Bandar. “FAST: A fuzzy semantic sentence similarity measure.” In Fuzzy Systems (FUZZ), 2013 IEEE International Conference on, pp. 1-8. IEEE, 2013.
6. M. Chong, L.Specia, and R. Mitkov, “Using natural language processing for automatic detection of plagiarism,” in in: Proceedings of the 4th International Plagiarism Conference (IPC-2010), 2010.
7. M. Mozgovoy, T. Kakkonen, and E. Sutinen, “Using natural language parsers in plagiarism detection,” in in: Proceedings of SLaTE’07 Workshop, 2007
8. Alberto Barrón-Cedeño, Paolo Rosso, José-Miguel Benedí “Reducing the plagiarism detection search space on the basis of the kullback-leibler distance,” in Proceedings of 10th International Conference on Computational Linguistics and Intelligent Text Processing, 2009, pp. 523–534.
9. J. Lewis, S. Ossowski, J. Hicks, M. Errami, and H. Garner, “Text similarity: An alternative way to search medline,” Bioinformatics, vol. 18, no. 22, pp. 2298–2304, 2006.
10. M. Errami, J. Hicks, W. Fisher, J. W. D. Trusty, T. Long, and H.Garner, “vu - a study of duplicate citations in medline,” Bioinformatics, vol. 24, pp. 243–249, 2008.

