Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 2, No. 1, February : 2023

GUIDELINES FOR FAIRLY MATCHED PARAGRAPHS IN RESEARCH PAPER

Rupali Chandrakar, Dr.Anurag Sharma, Prashant Kumar Tamrakar, Manisha Sahu, Sherin Koushor, Sanjay Rungta Group of Institutions, Bhilai, Chhattisgarh, India
rupalichandrakar226@gmail.com, anusiraag@gmail.com, prashant.tamrakar35@gmail.com, manishasahu0212@gmail.com, skoushor@rediffmail.com

Abstract
In any kind of research area, plagiarism play an very important role it is assumed by default is an crime or destruction of official document or patent. But in research point of view previous data have be borrowed by researcher. This point should be notified if the intention of a researcher was fair and if they were not aware about the plagiarism, so shall we call this as “Plagiarism”. Research papers can contain a variety of language modifications, including paraphrasing, summarising, semantic similarity checks, and concept borrowing. Even if all text modifications needed to be properly cited. In this paper we are trying to find out the matching among paragraphs (of suspicious manuscripts) and any paragraphs (of reference/source manuscripts) on any scientific research paper. The right guidelines and rules for paraphrasing plagiarism are covered in this paper also.

Keywords: paraphrasing, plagiarism policies, idea plagiarism, text reuse.

1. INTRODUCTION

Because of how simple information exchange is made possible by networking technologies, more literate people are looking for free text formats online. This facility has had a bad side effect where individuals are trying to author their work (writing, prepared and organised files, reports, scripts) by copying the concepts or ideas of others without giving due credit, especially in academia. Truly, section matching is a confounded interaction that requires a great deal of things to be settled. Some of them are fairly unimportant; some require a refined calculations and experimental guidelines to be applied to make copyright infringement location effective. For the most part essayist imagines that thought is free and we can take thought from anyplace however when a writer presents a thought, examines an issue, and offers an answer in a new or one of a kind way. It is the creators licensed innovation and should be refered to properly, for example with inside reference and outer reference on your works refered to page. It's vital to know the distinction between thoughts that are special to a creator and thoughts that are generally known and acknowledged, or in the public space. Data in the public space needn't bother with to be referred to.

2. THE STATE-OF-ART

Higher education bodies from every geographic region continue to deal plagiarism in best possible ways, all trying to create awareness on plagiarism among the academicians, scholars and writers by organizing awareness programs, conferences and workshops at national and international for making plagiarism policies, but all these policies and guidelines never land up in one frame till date. A noteworthy point of transition in policy timeline of plagiarism detection was prominently highlighted by Roig in 2013 who proposed a thorough arrangement of rules to moral writing practices that presents itself as a perceived instructional material supported by Office of Research Integrity (ORI) (M. Roig, 2013).

Yet another raised level of paraphrasing infringement was known as Mosaic/Hybrid/patchwork paraphrasing: This form of textual plagiarism is generally exhibited in suspicious documents by changing grammatical structures and narration forms of sentences forming the manuscripts by tactfully manipulating text with antonyms, translations, summarizations and also replacing selected phrases

with synonyms, homonyms and hyponyms that reflect the similar meaning to the original text (Choudhary et al., 2018), (Vani et al., 2016).
Like any other geographic domain, Institutional Academic Integrity Panel (IAIP) of India, too, has set guideline penalties against illegal plagiarized academic writing practices; the guidelines published since year 2017 (UGC regulations, 2017). The guidelines recommend 10% of document similarity as acceptable threshold norm for declaring the scholarly work to be fair and original one. Yet, the governance body does not recommend any standard plagiarism detection tool, complaint to penalty rules handling severity levels of plagiarism. This was the very objective of the put forth submitted work by the research scholar, who has attempted to keep the above-mentioned threshold norm, even in identifying various types and intensity levels of plagiarism within a suspicious (test) manuscript.
(Alzahrani et al. 2011) and (Vani et al, 2017) have pioneered the use of structural features for plagiarism detection, recently in this decade. Both these work groups favoured to the opinion that only summarized forms of text fragments, can be captured as idea plagiarized instances. However, both these works had been accomplished with the help of artificially generated corpora by mapping corresponding section and sub-section headings at various structural levels of document or paragraph representation. Unlike Alzahrani's way of formulating structural feature hierarchy into block-specific and content-specific; the scholar declares a holistic way of perceiving the textual features for accomplishing feature extraction step of plagiarism detection i.e. by parallel taking into account, two components of textual features namely, content specific and structure-specific features. Lexical, syntactic and semantic features together, were treated as content-specific features, while Alzahrani's multi-level document-paragraph-sentence representation formats, rightly justify the structure-specific features to compute similarity of a topic related content between sentential pairs or paragraph pairs or document pairs of suspicious and source manuscripts. (Pandey et al, 2018) have worked on compound evaluation metric which had two portions: semantic similarity by using relational similarity metric augmented by grammatical similarity for short and long lengths sentence pairs. Extending pandey's work of computing evaluation metric for paragraph pair for scientific paper (Pandey et al, 2019) reported a sentence similarity measure in multiple steps by representing the pair of sentences as joint noun phrases and joint verb phrases.

3. METHODOLOGY AND EXPERIMENTAL SETUP

The methodology has been developed to check the degree of similarity of each paragraphs of the suspicious manuscript with the each paragraphs of cited references manuscripts.
Take a research paper which is to be assumed as a suspicious manuscript and their all references manuscripts which is assumed to be source documents. Input the abstract of suspicious manuscript. Extract all NPs (Noun Phrases) from abstract of suspicious manuscript to form a vocabulary of document (most important words of any document), named it (abstract_vocab.txt file). Append keywords of suspicious manuscript with abstract _vocab.txt file to form seed_vocab.txt file.
abstract_vacab.txt + key_vocab.txt = seed_vacab.txt
Apply permutation and combination to make filter multiple strings matching from NPs & Keywords to filter most relevant source document. Search these NPs on each line of manuscript section wise and spot that line, any no of NPs may be found. Now search same NPs on each line of ref1, ref1.etc and find out NPs spotted line from references. Matching of the multiple phrases, contained in each line of manuscript doc with subset/superset/exact combination of those phrases in lines of references doc. In Stanford parser there are so many tags corresponding to the sentence structure and grammatical placements in the sentence. Such tags part sentence into noun, pronoun, verb, helping verb, auxiliary verb. In a sentence the noun family, verb family will always play an important role. So in the relation generation it is given more emphasis. All the noun forms and verbs forms are considered in the relationship. The noun family NN, NNS, NNP, and NNPS is considered. Now take any matched pair of sentences from manuscript and references and perform parsing on both sentences. Again extract the noun from that pair of sentences and match both sentences. If the percent of matching is less than 50%

then idea similarity found. Now type of sentences rewritten found, for deciding the sentence is plagiarised or not plagiarised, have to follow plagiarism guideline Rules. Before that we have to extract abstract and keyword section from any manuscript (Research paper).below is the small example of idea plagiarism found in research paper.

Manuscript Paragraph: Multi criteria decision making is a method to deal with the process of making decision among number of alternatives with conflicting criteria on them.
Reference Paragraph: Teachers are having many conflicting criteria among them and hence very difficult to decide their ranking, this will lead to multi criteria decision making.

Above example shows that manuscripts paragraph is plagiarised with reference paragraph and here no citation is done because of lack of plagiarism knowledge of author. So this type of plagiarism is easily detected by our experiments.

4. PARAGRAPH MATCHING GUIDELINES

As a result, there was a need to arrive at some precise set of plagiarism detection heuristics by conjoining conclusions from both the literature sources. The proposed criteria listed below are not fixed or valid for classifying plagiarism in every context. However applying criteria should help to be successful in the context of plagiarism detection test. These Rules are not mandatory; it may be modified according to situation. Here we only discuss the rule for paraphrasing plagiarism:

If (whole content or at least one content borrowed from the other original source given in references document AND (In-text / Reference citation missing) THEN this can be presumed as paraphrasing plagiarism.

Table 1: Fairly Paraphrasing Guideline Rules

<table>
<tr><th colspan="3">Do you carry idea from any other documents?</th></tr>
<tr><th colspan="2">Yes</th><th>No</th></tr>
<tr><td colspan="2">Is the carried content length from reference documents (Minimum 7-10 words continued)</td><td rowspan="2">May be general information or the writer's own ideas or Self Published work</td></tr>
<tr><td>Yes

Copy direct word -to-word & missing (quotation marks ∧ full in-text citation∧full references)</td><td>No
Paragraph idea & missing (in-text citation ∧ full references</td></tr>
<tr><td>Word To Word Plagiarism</td><td>Para Phrasing Plagiarism</td><td>Not plagiarism:</td></tr>
</table>

5. CONCLUSION:

In this paper, our goal was to check the authenticity of novel NLP approach based on the concept of paraphrasing plagiarism, used for identifying plagiarised level of suspicious documents in our research experiments .Nevertheless, the above set of experiments was repeated with slight different settings of applying Jaccard's similarity metric. As well as standard plagiarism policies for paraphrasing plagiarism is designed for the convenience of researcher and domain expert. In future we will try to implement these policies for other types of fairly text reusing act on research field.

Reference:

ACM Policy on Plagiarism, Misrepresentation, and Falsification: https://www.acm.org/publications/policies/plagiarism-overview
Ali, A. A., Hussam and Snasel, V. (2011).Overview and Comparison of Plagiarism Detection Tools.161-172.
Alzahrani, S. and Salim, N. 2010.Fuzzy semantic-based string similarity for extrinsic plagiarism detection. In Braschler, M., Harman, D. and Pianta, E. Eds, Lab Report for PAN at CLEF, 22-23 September, Vol. 1176, Padua, CEUR-WS.org.
Alzahrani, S., Salim, N. Abraham, A. and Palade, V. 2011.iPlag: intelligent plagiarism reasoner in scientific publications. In World Congress on Information and Communication Technologies WICT, IEEE, Los Alamitos, CA, 11-14 D.
Alzahrani, S., Salmon, M. and Abraham A. 2012.An Understanding plagiarism linguistic patterns, textual features, and detection methods. In IEEE transactions on systems, man, and cybernetics part c: application and reviews: 42 2 pp. 133-149.
Alzahrani, S., Salim N, and Palade, V. 2015.Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model. In Journal of King Saud University – Computer and Information Sciences 27, 248–268.
Angry.R, A., and Suharjito. 2014. Plagiarism detection algorithm using natural language processing based on grammar Analyzing. Journal of Theoretical and Applied Information Technology, Vol. 63 No.1
Miguel Roig 2006, Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing.
Canhasi, E., 2013. Measuring the sentence level similarity, ISCIM 2013, pp. 35-42 © 2013
Chong M., Special L. and Mitkov, R. 2013 Using Natural Language Processing for Automatic Detection of Plagiarism.
Chowdhury, H.A and Bhattacharyya, D.K. 2018. Plagiarism: Taxonomy, Tools and Detection Techniques.
Das D. and Pandey, S. 2015. A Survey Paper on in Text Citation of Manuscript with Computational Linguistic Tools. Journal of Emerging Technologies and Innovative Research (JETIR, Volume 2, Issue 3 JETIR (ISSN-2349-5162)
Lee, M.C., 2011. A novel sentence similarity measure for semantic based expert systems. Expert Syst. Appl. 38, 6392–6399.
Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., and Crockett, K., 2006. Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowledge Data Eng. 18, 1138–1150.
Ministry of human resource and development department, Government of India http://www.copyright.gov.in/Documents/Copy-Right-Rules-2012.pdf
Miranda Chong (2013) A Study on Plagiarism Detection and Plagiarism Direction Identification Using Natural Language Processing Techniques. PhD Thesis, University of Wolverhampton, UK (http://clg.wlv.ac.uk/papers/chong-thesis.pdf)
Miranda Chong, Lucia Specia, and Ruslan Mitkov.: Using Natural Language Processing for Automatic Detection of Plagiarism (2013).
Rao, S., Gupta, P., Singhal, K., and Majumder P. 2011. External &Intrinsic Plagiarism Detection: VSM & Discourse Markers based Approach. Notebook for PAN a CLEF.
Roig, M. 2015. Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing.
Roig, M., 2006.Avoiding plagiarism, self-plagiarism, and other questionable writing practices: a guide to ethical writing, in, St. Johns University.
Pandey, S., Sharma, H.R., & Rawal, A. 2016.A review paper on awareness statistics on plagiarism among research scholars. 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology RTEICT, 24-26.

S Pandey, A Rawal, An improved NLP approach for detection of plagiarism in scientific paper , International Journal of Applied Engineering Research, 2018.
Vani, K., Gupta, D. 2016. A Study on Extrinsic Text Plagiarism Detection Techniques and Tools.Journal of engineering science and technology review. 9. 150-164. 10.25103/jestr.094.23.

WEBLIOGRAPHY

1. http://www.webis.de
2. http://www.uni-weimar.de/en/universitty
3. http://www.ippheae.edu/
4. http://www.plagiraism.org/learning_center/typws_of_plagiarism.html
5. http://ww.wikipedia.org/wiki/intellectual_property
6. http://www.copyright.gov.in/documents/copyrightrules1957.pdf
7. http://www.copyright.gov.in/Documents/Copy-Right-Rules-2013.pdf
8. http:/www./mhrd.gov.in
9. https://www.programiz.com
10. https://www.codeproject.com/
11. https://stackoverflow.com
12. http://www.csse.monash.edu.au/projects/plague
13. http://www.cs.su.oz.au/~michaelw/YAP.html
14. https://jplag.ipd.kit.edu/
15. https://theory.stanford.edu/~aiken/moss/
16. http:// www.code-match.in
17. http://www.cshe.unimelb.edu.au/assessinglearning