International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 9 (2018) pp. 6733-6737

An Improved NLP Approach for Detection of Plagiarism in Scientific Paper

Shikha Pandey[1]

[1]*Department of Computer Science and Engineering, Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India.*

Dr. Arpana Rawal[2]

[2]*Department of Computer Science and Engineering, Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India.*

Abstract

The idea of doing things in a new way actually pressurizes the researcher to submit the composed distortion. At this time plagiarism takes place in the mind if researcher. Plagiarism detection as it is conceivable today does not ensure to detect plagiarism in all aspects; they can just distinguish duplicating, or all the more particularly comparable phrases. The proposed work means to outline fitting coordination for plagiarism detection upon scientific manuscript. In this investigation, we propose a creative NLP approach for intelligent plagiarism detection over one commercial tool. The plagiarism detection result contains both substance and area of re positions frame where the content pieces are acquired, in order to encourage simple human judgments.

Keywords: Plagiarism detection, NLP, Type dependency relationship, plagiarism taxonomy, idea plagiarism.

INTRODUCTION

In the period of World Wide Web, an ever increasing number of archives are being digitized and made easily accessible. Searching any information has become easier with a single click because variety of search engines and huge amount of online databases available. These favorable circumstances make the task of protecting intellectual property from information misuse become more troublesome. One of those dishonest behaviors is plagiarism; plagiarism causes significant damages on innovations. Most cases are identifying in academic works such as student assignments and research.

In all actuality Plagiarism Detection is an irritating procedure that requires a ton of things to be understood. Some of them are fairly significant; some require a modern calculations and exact view to be connected with a specific end goal to make plagiarism detection task proficient. The present work attempts to build up an intelligent semantic model for detecting similarity in free or organized texts by using advanced natural language processing techniques that is type dependency relationships on sentences.

LITERATURE SURVEY

The majority of the examination work has been seen on universal level. The majority of the scientists utilize Natural language identification procedure, this strategy showed up in the 1990s, and has delivered diverse location approaches.The most mainstream referred to works were distributed in the previous decade. Kong Leilei et al (2012) proposed two stage techniques for the tasks of Candidate Document Retrieval and Detailed Comparison for PAN 2012 challenge and scored first place for detailed comparison task. Candidate Document Retrieval have three stages Getting Query, Retrieving and Getting Sources for Plagiarized Passages, and in detailed comparison task have pre-handling, point by point correlation and post-preparing stages. The framework's general execution, particularly the review is higher than the vast majority of alternate techniques for most sorts of copyright infringement cases. The literary theft discovery technique they proposed is adaptable and versatile, as far as possible within time limit.

On the other hand, the pioneer work done by Alzahrani and group, they were able to distinguish the boundary conditions among varied plagiarism forms lately in the current decade (Salha Alzahrani et al 2011 - 2015). Towards this path, an attempt was made to develop an intelligent plagiarism reasoner tool: *iPlag*. This tool uses advanced methodology for exploiting document structure and performing n-gram based statistical overlap over similar document components ranked in order of statistical relevance (Salha Alzahrani et al, 2011). The tool generates output report for plagiarism along with plagiarism type remarks too.

Gabriel oberreuter et al (2011), proposed a method for external plagiarism, firstly identifies those pair of documents that have some text in common by using plagiarism search space reduction method, and then applied advance search to find plagiarized passages.

Du Zoo, et al (2010), South China University of Technology look into aggregate proposed a Cluster-Based Plagiarism Detection method which is one of the data recovery methods that were utilized as a part of numerous fields, such as text summarization, text classification and plagiarism detection. Its usability is to reduce the time of searching and comparison. This method scored 2nd rank in PAN-10 competition.

Kucecka T. (2011), which categorized the four types of documents obfuscation viz. cosmetic obfuscation, paraphrasing, scrambling obfuscation and snuffing obfuscation never the less, these unethical practices exhibit one or more combinations in literal or paraphrasing categories, as quoted in Roig's guidelines (M. Roig, 2006) (Kucecka T.,2011).

Efstathios stamatatos (2011), University of the Aegean, Greece research group proposed a method Based on Structural Information for plagiarism Detection to find out the Syntactic structure similarity of sentence with uses of n-grams approach.

It has been observed that surveyed literature till date demonstrate that until a decade ago, the majority of research was centered only on literal plagiarism like word-to-word plagiarism and paraphrasing plagiarism as reveled by plagiarism detection communities (Angry R.A. , 2014) and (S. Alzahrani, 2012). Non literal or intelligent plagiarism is still an untouched task for plagiarism detection tool maker as mentioned in a work by (S. Alzahrani, 2015). Today's plagiarist are very smart and they know that they will be catch any way by currently available tools, so they committing plagiarism in an intelligent way. By assessment and investigation of some of plagiarism detection tools and found that with smartly changed content are not recognized by these online accessible devices (seo small tool, Plagiarism Checker, Turnitin) by (Vani K et al, 2016). Miranda Chong et al (2011), Research Group in Computational Linguistics, University of Wolverhampton, UK, proposed a method for Automatic Detection of Plagiarism Using Natural Language Processing and attempted to find out the synonymy, paraphrasing similarity in sentence by using text pre-processing technique at preliminary stages.

EXPERIMENTAL SETUP

In this study, we performed detailed experiment for analyzing our proposed methodology. In this ways we performed an offline plagiarism detection task at paragraph level on a set of manuscripts; all manuscripts were part of input text. The approach was intended to recognize the level of closeness of every entry of the suspicious composition with referred to references original copies. Most important point to be noted here is this manuscript is already online available on a journal after being examined by PD tool. Still we discovered plagiarism in this document with their reference documents.

Manuscript Segmentation and Pre Processing

Our trial setup takes up plagiarism detection task from the cited set of references papers as the source articles. The detail of input manuscripts which is considered to be suspicious and reference manuscripts is shown in table 4.1. All reference manuscripts are the reference papers of suspicious manuscript as mentioned in reference section.

All manuscripts are segmented into the 3-4 lines of paragraphs before processing. The all suspicious paragraph composition is checked with each paragraphs of three reference article.

Table 3.1: Input of detail manuscript Summary for experiment

Suspicious Manuscript Title	**Source/Reference Manuscript Title**
"Statistical Survey on Big Data Analytics", (K. Prasad et al 2016)	1. "Data, DIKW, Big data and Data science". (Gu Jifaet al 2014) 2. "Beyond the hype: Big data concepts, methods, and analytics".(G. Amir et al, 2015) 3. "Applications of big Data: Current Status and Future Scope".(S.S. Karla, 2014)

Methodology of plagiarism detection

Now our steps for analysis of methodology are as follows:

1. Noun phrases (NPs) Extraction from the abstract portion of the suspicious manuscript and append keywords on it to form seed vocabulary.
2. Spotting the paragraph which contains at least one seed words. In our experiment the spotted paragraph from reference source is shown in below table 3.2.

Table 3.2: Details of Matched spotted paragraph.

Sl No.	Manuscripts	Total no of Matched Paragraphs
1	Suspicious	17
2	Ref1	9
3	Ref2	27
4	Ref3	14

Match at

1. Match all the paragraphs of suspicious manuscript with all the spotted paragraphs of reference manuscript with at least one commonly seed word. In this way, a finite set of reference paragraphs-pair combinations from suspicious manuscript and reference manuscripts are fetched for subsequent processing. A most extreme of 850 matching combination can be generated, but in this case 126 combinations for subsequent processing is revealed.

   Table3.3 shows the detail of matching combination.

Table 3.3: Manuscript*Reference paragraphs pair Matching

Manuscript*Reference paragraphs pair Matching		
1	17*9 = 153	57
2	17*27= 459	52
3	17*14=238	17
4	Maximum 850	126

2. For computing the similarity of two paragraph pair which is fetched in step 3, extract the parts of speech (POS) in the form of noun family and verb family to compute the similarity of two paragraphs (A, B). Expression (1) is used to calculate the similarity metric between paragraphs A & B symbolized by sim(A, B), which is thought to be at least 15% (assumed threshold).

$$sim(A, B) = S(A) \cap S(B) / S(A) \cup S(B) \quad (1)$$

Here, sim(A, B) shows common words appear from both

paragraphs, where S(A) and S(B) is a set of NPs and VPs family of words found in the paragraph A and B.

3. Here according to their percentage of similarity calculated from expression (1), we categorized the plagiarism type (which is not fixed, may be varying according to situation). Table shows the plagiarism type decided for this experiment.

Table 3.4: Details of Plagiarism Taxonomy.

Sl No.	Percentage found in expression (1)	Plagiarism Taxonomy
1	Between 15% to 30%	Go to next step
2	Between 31% to 50%	Paraphrasing or Semantic
3	Between 51% to 70%	Paraphrasing
4	Greater than 71%	Word-to-Word

4. The last step is calculated for idea similarity only when the paragraph pairs skip the condition of similarity computation labeled as sim $_{(synonym)}$ similarity metric. In this step, typed dependency structures of sentential pairs are checked for semantic similarity denoted by sim_{rel} (A, B), such that,

$$sim_{rel}(A,B)=\frac{1^{*}\left|S(A,n\text{Ç}S(B,n)\right|+0.67^{*}\left|S(A,n)\text{Ç}S(B,n)\right|+0.33^{*}\left|S(A,n)\text{Ç}S(B,n)\right|}{Min\,(countA,countB)} \quad (2)$$

In expression (2), S (A) and S (B) represent the typed dependency relationships of paragraph pairs. Also, the three weighted terms of numerator component indicate the different degrees of semantic overlap i.e. complete, partial and minimum overlaps, observed upon the matched typed-dependency relations extracted out of the input paragraph pairs. As a result of the above two steps of similarity computation, the similarity values were tabulated for suspicious (test) manuscript against three reference manuscripts in table 3.5.

Table 3.5: Comparison between Suspicious as well as reference 1, reference 2 and reference 3 manuscripts

Sl. No.	Suspicious Manuscript	Reference Manuscripts	Similarity (Exp1)	Similarity (Exp 2)	Type of plagiarism
1	<M>PID1	<R1>PID7	22.22	35.22	Idea
2	<M>PID2	<R1>PID8	34.29	-------	paraphrasig
3	<M>PID3	<R2>PID10	39.29	------	paraphrasing
4	<M>PID10	<R2>PID17	30.61	-----	paraphrasing
5	<M>PID11	<R2>ParaI21	20.29	56.27	Idea
6	<M>PID12	<R2>PID22	22.45	56.31	Idea
7	<M>PID12	<R2>PID23	16.95	53.14	Idea
8	<M>PID13	<R2>PID24	36.36	----	paraphrasing
9	<M>PID14	<R2>PID25	52.86	----	Paraphrasing
10	<M>PID17	<R2>PID5	18.92	75.55	Idea
11	<M>PID17	<R2>PID7	17.24	45.61	Idea
12	<M>PID17	<R2>PID18	19.05	58.25	Idea
13	<M>PID17	<R2>PID26	50.00	-----	paraphrasing
14	<M>PID8	<R3>PID6	18.18	33.40	Idea
15	<M>PID16	<R3>PID14	25.0	69.33	Idea
16	<M>PID17	<R3>PID2	15.79	24.36	Idea

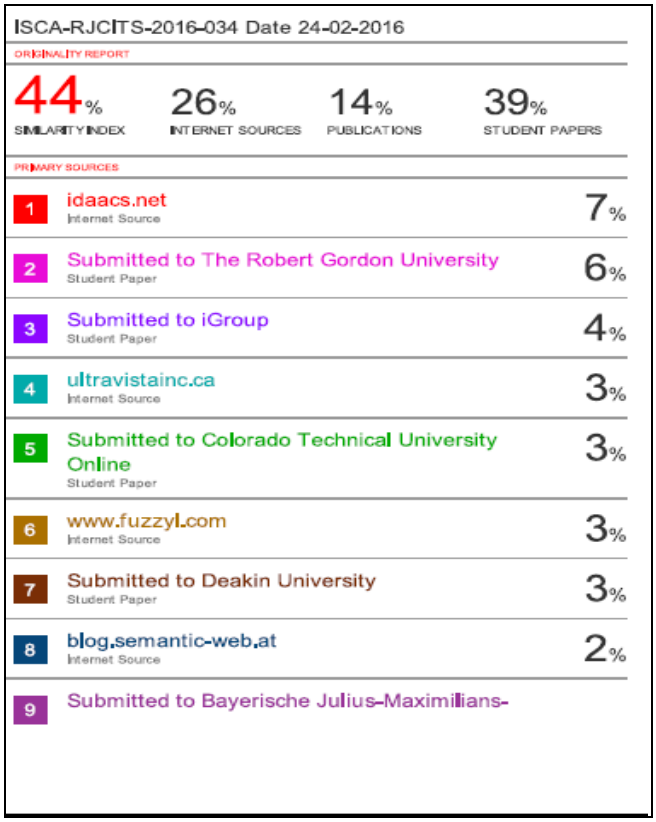

Figure 3.1 Plagiarism Report generated by plagiarism-checker tool, (Courtesy: ISCA - RJCITS-2016)

Performance Assessment and Observations

Figure 3.1, shows the plagiarism report generated by plagiarism detection tools used by ISCA-RJCIT-2016 dated on 24-02-2016 and Table 3.5, shows our in-depth analysis of plagiarism report with plagiarism type. Our exploratory outcomes investigate each section of the suspicious manuscript with the candidate manuscripts in spite of this PLAGIARISM DETECTION device delivered just outlined outcome.

The detailed comparison of both results is as follows:

1. The time complexity is good because the proposed framework did not performed text pre-processing techniques.
2. Our precision and recall rate is additionally great because it works offline and our sources are the cited references of suspicious manuscript, available PD tool takes excessively time since it looks universally to search right source for comparison.
3. From point by point examination we have discovered that suspicious is all most replicated from Reference1, Reference2 and Reference3.
4. All most basic closeness with reference2 has been observed, however ISCA-RJCITS Plagiarism detection tool did not perceives any similitude with Reference2 in spite of these demonstrating likeness with different other materials which is not relevant to source.

CONCLUSION, SOCIAL BENEFITS OF PROPOSED SYSTEM AND FUTURE SCOPE

Currently available PD tool are lacking to define the plagiarism type only generates the source of plagiarism. In this study we performed plagiarism detection task and generated the report which labeled with type of plagiarism although the generated report is in tabular form but the current work is in progress and tries to make plagiarism report in presentable form.

In the event that researcher with fair intentions duplicating and sticking to other work or form their own work, inquire about paper and draw references from other sources yet regularly overlook referring to the source properly. So our framework gives assistance to cross verification for reference. For Commentator: - It is useful for the analyst to distinguish written falsification of research paper with moral points of confinement and given due dates on the grounds that our framework produces come about with legitimate rules. For College/university: - If the college or university makes some legitimate approaches for written falsification so this product is exceptionally helpful for making rules and standards according to plagiarism type and it will be open and clear for understudy who confer unoriginality. From the above exploratory outcome it has been observed manually that figure 1, 2, 4 in suspicious manuscript is also straight forwardly duplicated from Reference2. However, PD s/w cannot distinguish the figures, tables and charts etc, it only works on text. Our framework is likewise not ready to recognize these. So future work incorporate attempt to identification counterfeiting on images, tables and charts. The proposed models also intends to make the machine assisted plagiarism policy/guidelines to better distinguish the plagiarism heuristics which does not exist in the current scenario.

REFERENCES

[1] Alzahrani, S. M., Naomie Salim, and Vasile Palade: Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model, Journal of King Saud University – Computer and Information Sciences 27, 248–268(2015).

[2] Alzahrani, S. M., Salim, N., Abraham, and A.: Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE transactions on systems, man, and cybernetics part c: application and reviews, 42 (2) (2012).

[3] Alzahrani, S. M.: iPlag: Intelligent Plagiarism Reasoner in Scientific Publications, IEEE (2011).

[4] Angry Ronald Adam et al, "plagiarism detection algorithm using natural language processing based on grammar Analyzing", Journal of Theoretical and Applied Information Technology 10th May 2014. Vol. 63 No.1

[5] Christina Kraus: Plagiarism Detection-State-of-the-art systems (2016) and evaluation method (2016).

[6] Du Zoo, Wei-kiang Long, Zhang Ling: A Cluster-Based Plagiarism Detection Method - Lab Report for PAN at CLEF 2010. CLEF (Notebook Papers/LABs/Workshops) 2010.

[7] Efstathios Stamatatos" Plagiarism Detection Based on Structural Information" University of the Aegean, Greece, 2011.

[8] Gabriel Oberreuter, Gaston L'Huillier, Sebastián A. Ríos, and Juan D. Velásquez, " Approaches for Intrinsic and External Plagiarism Detection " Notebook for PAN at CLEF 2011.

[9] Gu Jifa, Zhang Lingling: Data, DIKW, Big data and Data science. 2nd International Conference on Information Technology and Quantitative Management, ITQM, *Procedia Computer Science,* 31, 814-821(2014).

[10] Gandomi Amir, Haider Murtaza.: Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35, 137-14 (2015)

[11] Indian University Bloomington, http://www/indiana.edu-(2017)

[12] Kauleshwar Prasad, Arpana Rawal.: Statistical Survey on Big Data Analytics, Research Journal of Computer and Information Technology Sciences, Vol. 4(9), 22-24, September (2016).

[13] Kong Leilei1, Qi Haoliang1, Wang Shuai1, Du Cuixia2,Wang Suhong2 and Han Yong1 "Approaches for Candidate Document Retrieval and Detailed Comparison of Plagiarism Detection "Notebook for PAN at CLEF 2012.

[14] Sabia Sheetal Kalra: Applications of Big Data: Current Status and Future Scope. *International*

Journal on Advanced Computer Theory and Engineering, 25-29 (2014).

[15] Marie-Catherine de, Marneffe, Bill McCartney, Christopher D. Manning, In *LREC (2006)*.

[16] Miguel Roig.: Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing (2015).

[17] Miranda Chong*, Lucia Specia, Ruslan Mitkov.: Using Natural Language Processing for Automatic Detection of Plagiarism (2011).

[18] Vani K., Deepak Gupta.: Study on Extrinsic Text Plagiarism Detection Techniques and Tools, Journal of Engineering Science and Technology Review 9 (4) (2016).